Relatively Smart: A New Approach for Instance-Optimal Learning

Shaddin Dughmi[*] SHADDIN@USC.EDU
University of Southern California

Alireza F. Pour[†] ALIREZA.FATHOLLAHPOUR@UWATERLOO.CA
University of Waterloo

Editors: Steve Hanneke and Tor Lattimore

Abstract

We revisit the framework of *Smart PAC learning*, which seeks supervised learners which compete with semi-supervised learners that are provided full knowledge of the *marginal* distribution on unlabeled data. Prior work has shown that such marginal-by-marginal guarantees are possible for "most" marginals, with respect to an arbitrary fixed and known measure, but not more generally. We discover that this failure can be attributed to an "indistinguishability" phenomenon: There are marginals which cannot be statistically distinguished from other marginals that require different learning approaches. In such settings, semi-supervised learning cannot certify its guarantees from unlabeled data, rendering them arguably non-actionable.

We propose *relatively smart learning*, a new framework which demands that a supervised learner compete only with the best "certifiable" semi-supervised guarantee. We show that such modest relaxation suffices to bypass the impossibility results from prior work. In the distribution-free setting, we show that the One-Inclusion Graph learner is relatively smart up to squaring the sample complexity, and show that no supervised learning algorithm can do better. For distribution-family settings, we show that relatively smart learning can be impossible or can require idiosyncratic learning approaches, and its difficulty can be non-monotone in the inclusion order on distribution families.

Keywords: Smart Learning, Semi-Supervised Learning, Distribution-Fixed Learning, One-Inclusion Graph Algorithm

1. Introduction

A substantial portion of research in learning theory proceeds within the worst-case tradition characteristic of theoretical computer science more broadly. This includes, most prominently, the PAC model (Valiant, 1984) and its numerous extensions, which typically impose a "flat" inductive bias on the hypothesis class and/or data-generating distribution, then leave the choice of instance to an adversary. Performance guarantees are consequently evaluated in the worst case over all hypotheses and distributions consistent with these assumptions.

It has been argued that this perspective is somewhat removed from practical machine learning, which tends to be more adaptive to the specifics of its deployment domain. Techniques such as unsupervised pretraining, hyperparameter optimization, post-hoc model refinement, and incorporation of domain expertise are often used to tailor the learner to the data distribution. Partly as a result

[*] Supported by NSF Grant CCF-2432219. Part of this work was done while the author was on sabbatical as the Carter and Tania Neild visiting professor at Northwestern University, as well as a visiting professor in the Data Science Institute at the University of Chicago.
[†] Supported by a David Cheriton Scholarship and a Vector Institute Research Grant.

of these adaptation mechanisms, the observed performance of real-world machine learning systems is typically far from worst-case (Chapelle et al., 2006; Erhan et al., 2010; Guo et al., 2017; Zhang et al., 2017; Frankle and Carbin, 2019; Devlin et al., 2019; Chen et al., 2020; Yang and Shami, 2020).

In this paper, we investigate one facet of this divide: how a learner can be tailored to the unlabeled data distribution, also known as the *marginal distribution*. We focus on statistical aspects of this question, as captured by the sample complexity (or equivalently, the error rate) of learning. We restrict attention to what is arguably the most historically instructive learning setting: realizable binary classification.

Distribution-Fixed and Semi-Supervised Learning

The original PAC learning model considers *fully supervised learners* which only receive labeled samples from the distribution. A particularly powerful extension of this paradigm additionally provides the learner with full knowledge of the marginal distribution of unlabeled data. This *distribution-fixed* model of learning was studied by Benedek and Itai (1991), where they characterize learnability qualitatively for each marginal in terms of the existence of finite covers of the hypothesis class, at every error scale, with respect to the disagreement metric induced by the marginal. They also derive (not quite tight) upper and lower bounds on the sample complexity in terms of these covers. Formal evidence that knowing the marginal can enable learning was subsequently provided by Dudley et al. (1994).

As pointed out by Ben-David et al. (2008), distribution-fixed learning can be viewed as a utopian idealization of *semi-supervised learning*—the paradigm which augments labeled data with more plentiful unlabeled data. There is a long line of work which explores more realistic formulations of semi-supervised learning, seeking to understand whether, when, and how much finite unlabeled data helps in learning (e.g. Ben-David et al., 2008; Balcan and Blum, 2010; Darnstädt et al., 2013; Globerson et al., 2017; Göpfert et al., 2019; Golovnev et al., 2019; Pukdee et al., 2023). There are many shades of these questions explored by this body of work to which we cannot do justice, but for our purposes the gist is as follows: Unlabeled data does not improve minimax error rates in distribution-free settings,[1] but can lead to drastic marginal-by-marginal improvements, as well as minimax improvements for some distribution-family settings and under "compatibility" assumptions between marginals and hypotheses.

Smart Learning

Instead of further exploring the power and limits of semi-supervised learning, we take a different tack in this paper. We build on the closely-related framework of *Smart Learning* introduced by Darnstädt and Simon (2011). Roughly speaking, a smart learner is a fully-supervised learner which *does about as well as if it knew the marginal distribution already, even though it doesn't.* This is an (approximate) instance-optimality guarantee with respect to marginals: a smart learner approximately matches the optimal distribution-fixed error rate (or equivalently, sample complexity) for every marginal simultaneously.

While ambitious, the goal of smart learning seems plausible at first glance: By eschewing minimax guarantees across marginals, the learner is permitted to perform poorly for "hard" marginals such as those appearing in the fundamental theorem of PAC learning, while paying special attention to those marginals most amenable to semi-supervised learning techniques. Indeed, Darnstädt and

1. Recall that the lower-bound in the fundamental theorem of PAC learning is robust to knowledge of the marginal.

Simon (2011) show a compelling, albeit qualified, positive result via an innovative application of the minimax theorem for zero-sum games: Smart learning is possible in general distribution-free settings for "most" marginals, where "most" is quantified with respect a prior distribution on marginals that is given in advance. Unfortunately, any hope of removing this qualification was dashed by subsequent work of Darnstädt et al. (2013), strengthening an earlier result of Dudley et al. (1994): They exhibit a hypothesis class and family of marginals with distribution-fixed error rates rapidly and uniformly tending to zero, whereas for any fully-supervised learner—not equipped with foreknowledge of the marginal—and any finite sample size there is an instance where the learner performs essentially no better than random guessing.

From Smart to Relatively Smart Learning

The results of (Darnstädt and Simon, 2011; Darnstädt et al., 2013) might appear to close the book on smart learning for general hypothesis classes and sufficiently rich distribution families. This, however, is where our work comes in. Our contribution emanates from the following realization: Smart learning fails when a distribution-fixed learner $\mathcal{A}_{\mathcal{D}}$ catered to a marginal $\mathcal{D}$ cannot use its unlabeled data to distinguish $\mathcal{D}$ from other marginals $\mathcal{D}'$ where $\mathcal{A}_{\mathcal{D}}$ performs much worse. In other words, it is impossible to detect misspecifications of the marginal that are consequential to learning by merely inspecting the unlabeled data, rendering error guarantees impossible to certify prior to procuring labels. Our results will imply that this is the only qualitative obstacle to smart learning.[2]

Sparked by this realization we introduce *relatively smart learning*, which minimally relaxes smart learning to "price-in" the above-described obstacle for each marginal distribution $\mathcal{D}$. Informally, for each marginal $\mathcal{D}$ and learner $\mathcal{A} = \mathcal{A}_{\mathcal{D}}$ we seek to compete not with the error $\mathcal{A}$ incurs on $\mathcal{D}$ (as in smart learning), but rather with the best *certifiable upperbound* on that error that can be calculated from the unlabeled data. By this we mean that there is a real-valued *certifier* $\mathcal{C}$ which estimates $\mathcal{A}$'s error from the unlabeled data, and we require $\mathcal{C}$ to be *sound* in the following strong sense: its error estimate must in-expectation upper bound $\mathcal{A}$'s error for all admissible instances. Importantly, even if $\mathcal{A} = \mathcal{A}_{\mathcal{D}}$ is tailored to a particular marginal $\mathcal{D}$, $\mathcal{C}$ must soundly certify $\mathcal{A}$'s error for all admissible marginals $\mathcal{D}'$, even if $\mathcal{D}' \neq \mathcal{D}$. This requirement serves to relax distribution-fixed error rates upwards by effectively taking the worst case over all $\mathcal{D}'$ that are indistinguishable from $\mathcal{D}$, making the design of instance-wise competitive learners more achievable. A fully-supervised learner is now said to be *relatively smart*[3] if it (approximately) matches the best certifiable error rate for every admissible marginal distribution.

Our Results

We begin in the distribution-free setting with general hypothesis classes. Our main positive result (Theorem 3.2) is that relatively smart learning is possible at the cost of a quadratic blowup in the number of samples and constant blowup in the error, as compared to the best certifiable distribution-dependent error rates. In particular, this is achieved by the familiar One-Inclusion-Graph (OIG) learner of Haussler et al. (1994). Our main negative result (Theorem 4.1) is that this is essentially tight, in that every relatively smart learner must suffer a near-quadratic blowup in sample complex-

2. Smart learning is self-evidently an unsupervised learning task: that of learning characteristics of an unknown $\mathcal{D}$ that are most pertinent for subsequent supervised learning. Our stated obstacle concerns a prima facie easier task more akin to testing: $\mathcal{D}$ is fixed and must merely be distinguished from other distributions $\mathcal{D}'$ which prohibit similar supervised learning approaches. Perhaps surprisingly, our results reveal an equivalence between learning and testing which holds here, but fails more generally in statistics (see e.g. Batu et al. (2000)).
3. Read: Smart relative to every certifiable error guarantee.

ity. Since the latter result is quite technical, we also provide a simpler proof of the same bound for the special case of the OIG and Empirical Risk Minimization (ERM) learners (Theorem 3.1). We also discuss the intriguing question of whether ERM, or some other "simple" and typically-tractable learner, is relatively smart (Open Question 3.3).

We then examine distribution-family settings. We observe in Corollary 5.1 that our main positive result (Theorem 3.2) extends to families characterized only by the allowable subsets of the domain on which data must be supported (e.g. manifolds satisfying some algebraic or topological requirements). Beyond such "simple" families, we show that relatively smart learning starts to exhibit richer and more nuanced structure: There are families where relatively smart learning is completely impossible (Theorem 5.3), and others where it is possible but not by straightforward approaches such as OIG or ERM (Theorem 5.2). On a more meta level, we show in Corollary 5.4 that the difficulty of relatively smart learning, unlike traditional PAC learning or Smart learning, can be non-monotone in the inclusion order on distribution families. We attribute this to the shifting benchmark of certifiable error rates, where the soundness requirement introduces dependence on the family as a whole.

We note that our negative results hold for countable domains, and our positive results hold more generally for domains, hypothesis classes, and distributions jointly satisfying standard measurability assumptions (see e.g. Shalev-Shwartz and Ben-David, 2014). As is common in learning theory, we take a hands-off approach to the measure-theoretic details.

Connection to Testable Learning

We would be remiss not to discuss connections between relatively smart learning and the framework of *testable learning*, originally introduced by Rubinfeld and Vasilyan (2023) and spawning a rapid succession of followup work since (Gollakota et al., 2023b,a; Diakonikolas et al., 2023; Klivans et al., 2024; Gollakota et al., 2024). Our certifiers can be viewed as real-valued analogues to the testers from that framework, where soundness in our framework is analogous to their requirement that the learner performs well for every distribution which passes the test. Whereas testable learning is concerned with the design of learner/tester pairs for a specific distribution or distributional property, we instead use these objects as our *benchmark* for *every distribution separately*. This is the essence of the connection, as well as the main difference.

There are other important differences: (a) The literature on testable learning is primarily concerned with computational complexity, with the notable exception of Gollakota et al. (2023b). (b) Testers in that framework are permitted to use labeled data, though this happens to not be necessary for many of the problems considered. (c) Their focus is on agnostic learning, with the analogous realizable question rendered trivial by (b); a notable exception is in the context of distribution shift (Klivans et al., 2024).

Roadmap

Section 2 introduces relatively smart learning and discusses its basic properties. Section 3 examines the familiar ERM and OIG learners in the context of distribution-free relatively smart learning, and presents our main positive result for OIG as well as a complementary negative result for both learners. Section 4 provides a tight negative result for distribution-free relatively smart learning which holds for all learners. Section 5 explores relatively-smart learning in distribution-family settings, outlining similarities and differences from the distribution-free setting.

Basic Notation

For a probability distribution $\mathcal{D}$ on some domain $\mathcal{X}$ and an event $Y \subseteq \mathcal{X}$, we use $\mathcal{D}[Y] = \mathbb{P}_{\mathcal{D}}[Y]$ as shorthand for the probability of Y, and $\mathcal{D}_{|Y} = \mathbb{P}_{\mathcal{D}}[\,.\,|Y]$ as shorthand for the conditional distribution of $\mathcal{D}$ given Y. When the domain $\mathcal{X}$ is countable we use $\mathcal{D}[x] = \mathbb{P}_{\mathcal{D}}[x]$ to denote the probability of $x \in \mathcal{X}$, and use $\mathsf{supp}(\mathcal{D}) = \{x \in \mathcal{X} : \mathcal{D}[x] > 0\}$ to denote the support of $\mathcal{D}$. For a finite multiset S we use $\mathcal{D}_S$ to denote the uniform distribution on S. Given a function h defined on some domain $\mathcal{X}$, we use $h|Y$ to denote its restriction to some $Y \subseteq \mathcal{X}$. For a set $\mathcal{X}$ we use $\mathcal{X}^*$ to denote the family of finite sequences over $\mathcal{X}$. For a predicate or probability event E we use $\mathbb{1}[E] \in \{0,1\}$ to denote the indicator of E. Finally, we use standard order-of-growth (big-Oh) notation, though for functions $f(\eta, m)$ and $g(m)$ we write $f = O_\eta(g)$ to indicate that $f = O(g)$ whenever η is a fixed constant.

2. Relatively Smart Learning

We begin in the standard PAC learning setting of realizable binary classification. There is a *data domain* $\mathcal{X}$ and a *hypothesis class* $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Unlabeled data comes from a *marginal distribution* $\mathcal{D}$ on $\mathcal{X}$, and is labeled with some *ground truth* hypothesis $h \in \mathcal{H}$. We use $\mathcal{D}_h$ to denote the joint distribution of labeled data (x, y) with $x \sim \mathcal{D}$ and $y = h(x)$. In the *distribution-free setting* $\mathcal{D}$ can be arbitrary; more generally, we also consider settings where $\mathcal{D}$ is restricted to some *distribution-family* $\mathbb{D}$. For a *predictor* $f : \mathcal{X} \to \{0,1\}$ and labeled data distribution $\mathcal{D}_h$, we denote the *expected 0-1 loss*, or simply *loss* or *error*, of f on $\mathcal{D}_h$ by $L(f, D_h) = \mathbb{E}_{(x,y)\sim\mathcal{D}_h}[\mathbb{1}[f(x) \neq y]]$. For a finite multiset T of labeled data we overload notation and denote $L(f, T) = L(f, \mathcal{D}_T) = \frac{1}{|T|}\sum_{(x,y)\in T} \mathbb{1}[f(x) \neq y]$.

A *(fully-supervised) learner* $\mathcal{A}$ takes as input a sequence $S = (x_1, y_1), \ldots, (x_m, y_m)$ of labeled samples—often referred to as *training data*—drawn i.i.d. from $\mathcal{D}_h$, and outputs a predictor $\mathcal{A}(S) : \mathcal{X} \to \{0,1\}$. We depart from the usual PAC approach by evaluating a learner by its expected error rather than its high probability error; this is merely for expository convenience, as all our results translate to PAC guarantees by standard arguments. More importantly, we measure a learner's error as a function of the marginal $\mathcal{D}$, which permits comparing learners catered to the marginal—we call these *$\mathcal{D}$-fixed learners*—to learners which are provided no such knowledge. We therefore define distribution-dependent error rates as follows.

Definition 2.1 *A* distribution-dependent error rate *is a function $\epsilon : \mathbb{D} \times \mathbb{N} \to [0,1]$, where $\epsilon(\mathcal{D}, m)$ is an error associated with a distribution $\mathcal{D} \in \mathbb{D}$ and a number of samples m.*

Definition 2.2 *For a learner $\mathcal{A}$, let its distribution-dependent error rate $\epsilon_{\mathcal{A}}(\mathcal{D}, m)$ be the worst-case, over hypotheses $h \in \mathcal{H}$, of its expected error with respect to $\mathcal{D}_h$ when given m i.i.d. samples from $\mathcal{D}_h$, i.e., $\epsilon_{\mathcal{A}}(\mathcal{D}, m) = \sup_{h\in\mathcal{H}} \mathbb{E}_{S\sim\mathcal{D}_h^m}[L(\mathcal{A}(S), \mathcal{D}_h)]$.*

We pair learners with functions that *certify* their distribution-dependent errors from unlabeled data. We refer to those as *certifiers*. We require certifiers to be *sound*, meaning that they never underestimate the learner's error for any distribution in the class.

Definition 2.3 (Sound certifier) *Let $\mathcal{A}$ be a learner. We say a function $\mathcal{C} : \mathcal{X}^* \to [0,1]$ is a* sound certifier *for $\mathcal{A}$ if for every $\mathcal{D} \in \mathbb{D}$, $m \in \mathbb{N}$, and $S \sim \mathcal{D}^m$ we have $\mathbb{E}[\mathcal{C}(S)] \geq \epsilon_{\mathcal{A}}(\mathcal{D}, m)$.*

Whereas smart PAC learning seeks to compete with the best distribution-fixed learner for each marginal, we instead propose a more modest benchmark: We only credit a distribution-fixed learner with error rates witnessed by a sound certifier. This gives rise to *certifiable error rates*.

Definition 2.4 (Certifiable error rate) *A distribution-dependent error rate $\epsilon(.,.)$ is* certifiable *if for each $\mathcal{D} \in \mathbb{D}$, there exists a $\mathcal{D}$-fixed learner $\mathcal{A}$ and a sound certifier $\mathcal{C}$ for $\mathcal{A}$ such that for each $m \in \mathbb{N}$ and $S \sim \mathcal{D}^m$, we have $\mathbb{E}[\mathcal{C}(S)] \leq \epsilon(\mathcal{D}, m)$.*

Note that in the above definition, we also have $\epsilon_{\mathcal{A}}(\mathcal{D}, m) \leq \epsilon(\mathcal{D}, m)$ for $\mathcal{A} = \mathcal{A}_{\mathcal{D}}$ by soundness of $\mathcal{C}$ for $\mathcal{A}$. More importantly, the bite in this definition comes from the fact that we require $\mathcal{C}$ to be sound for $\mathcal{A}$ everywhere (i.e., for all $\mathcal{D}' \in \mathbb{D}$), even though $\mathcal{A}$ is catered to $\mathcal{D}$ specifically. This is what effectively forces us to take the worst case error rate over all $\mathcal{D}'$ that are indistinguishable from $\mathcal{D}$ with m samples. It is also important to note that our definition allows the certifiable error rate to be achieved by different learners $\mathcal{A}_{\mathcal{D}}$ that are fixed for each distribution $\mathcal{D} \in \mathbb{D}$. The only requirement is that the learner has a certifier that can soundly witness its purported error rate for *all* distributions in $\mathbb{D}$, even those distributions on which the learner is not designed to perform well.

We next give a simple example illustrating when certification is possible and when it is not. This example also serves as a building block in Theorem 3.1 to prove our negative results for ERM and OIG. Consider a hypothesis class on $[n]$ where every hypothesis has all but $\sqrt{n}$ points labeled the same way; i.e., with at least $n - \sqrt{n}$ zeros or at least $n - \sqrt{n}$ ones. Under the uniform distribution on $[n]$, the *majority learner* which always predicts the most frequently seen label in training has expected error on the order of $1/\sqrt{n}$, even with a single labeled sample. This learner is, in a sense, catered to the uniform distribution: it may have large error on marginals that put much larger mass on points with the minority label. Certifying an error of $O(1/\sqrt{n})$ for the uniform distribution then hinges on whether such marginals can be distinguished from uniform using unlabeled samples. This is possible using techniques from *uniformity testing* precisely when the number of samples m is at least on the order of $\sqrt{n}$. In other words, though the majority learner achieves error rate on the order of $1\sqrt{n}$ for all $m \geq 1$, this is only certifiable for $m = \Omega(\sqrt{n})$. More generally, anytime a learner's error guarantee hinges on some property of the marginal distribution, the extent to which certification is possible depends on whether the property can be detected from unlabeled data.

We can now define relatively smart learning, a relaxation of smart learning which judges a learner relative to the best certifiable error rate for each distribution. Informally speaking, relatively smart learning lets us "off the hook" whenever small distribution-fixed errors cannot be certified from unlabeled data.

Definition 2.5 (Relatively Smart Learning) *For a function $\sigma : \mathbb{N} \times (0,1) \to \mathbb{N}$ and constant $\alpha > 0$, we call a learner $\mathcal{A}$* relatively (α, σ)-smart *if*

$$\epsilon_{\mathcal{A}}(\mathcal{D}, \sigma(m, \eta)) \leq \alpha \epsilon(\mathcal{D}, m) + \eta$$

for every certifiable distribution-dependent error rate ϵ, every distribution $\mathcal{D} \in \mathbb{D}$, every sample size $m \in \mathbb{N}$, and every additive error parameter $\eta \in (0,1)$. We also say $\mathcal{A}$ is relatively smart *if there exist α and σ such that it is relatively (α, σ)-smart.*

Note that we allow a relatively smart learner's error to trail certifiable errors by a constant multiplicative term α and an additive term η, so long as η can be made arbitrarily small.[4] We also allow the relatively smart learner to trail in the number of samples by an amount which can depend on η, as described by the *sample blowup function* $\sigma(m, \eta)$. One could parameterize smart learning

4. Arbitrarily-small additive error η features in our main positive result. It is an interesting and seemingly-challenging question whether fixing $\eta = 0$ permits relatively smart learning.

similarly by removing the terms "certifiable" and "relatively" from Definition 2.5, though the strong impossibility result of (Darnstädt et al., 2013, Theorem 2)—discussed in Section 1—persists even with the allowances provided by α, η, and σ.

One might initially hope to construct relatively smart learners with $\alpha = O(1)$ and $\sigma(m, \eta) = O(m)$ for each fixed $\eta > 0$ in fairly general settings. This turns out to be asking too much. The interesting question is therefore which, if any, sample blowup functions $\sigma(m, \eta)$ permit relatively smart learning.

3. Distribution-Free Setting: OIG and ERM

We begin our exploration of relatively smart learning in the distribution-free setting with the familiar *one-inclusion graph (OIG)* and *empirical risk minimization (ERM)* learners. Our most notable result here is that the OIG learner is relatively smart with only a quadratic blowup in sample complexity. We show that this is essentially tight by way of a quadratic lowerbound which holds for both the OIG and ERM learners. We leave wide open whether ERM is relatively smart, and discuss associated challenges.

We use standard definitions for ERM and OIG, which can be found in Appendix A. For purposes of arguments presented in this section, the reader need only keep in mind the following defining properties of the two learners:

- ERM outputs a hypothesis consistent with the (realizable) training data, with ties among such hypotheses broken adversarially.
- For each unlabeled dataset $S = (x_1, \ldots, x_n) \in \mathcal{X}^*$, the OIG learner minimizes the worst-case *transductive error*, also often known as *leave-one-out error*, among all learners. The worst-case transductive error of a learner $\mathcal{A}$ on unlabeled dataset S is defined as follows:

$$\epsilon_{\mathcal{A}}^{\mathsf{Trans}}(S) = \max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[\mathcal{A}\left(S_h^{(-i)}\right)(x_i) \neq h(x_i)\right],$$

  where $S_h = (x_1, h(x_1)), \ldots, (x_n, h(x_n))$ is the labeled dataset corresponding to S and $h \in \mathcal{H}$, and the training data $S_h^{(-i)}$ consists of S_h with its ith sample $(x_i, h(x_i))$ omitted.

We present our negative result for the OIG and ERM learners first. Though this is subsumed by the more general impossibility result in Section 4, and moreover technically modest, we find the associated arguments a suitable warmup for appreciating the nuances of relatively smart learning.

Theorem 3.1 *There exists a hypothesis class on a countable domain such that for every pair of constants α and β and every function $\sigma(m, \eta) = O_\eta(m^{2-\beta})$, the ERM and OIG learners fail to be relatively (α, σ)-smart in the distribution-free setting.*

Notably, the certifiable error rates with which ERM and OIG cannot compete are achieved by the simple *majority learner*, which just outputs the most frequent label in the training data.

The formal proof of Theorem 3.1 can be found in Appendix C, but the high-level idea is as follows. Consider the following hypothesis class on n datapoints: At least $n - o(n)$ datapoints have the same *majority label* (whether 0 or 1), with the remaining $o(n)$ points having the other *minority label*. Think of the number of minority labels as being only slightly sublinear, e.g. $n^{0.99}$. When faced with a uniform (or nearly uniform) distribution over the n points, the majority learner quickly

approaches a vanishing error rate $o(n)/n = o(1)$, even with very few samples m. A certifier can identify such a nearly uniform distribution on n points when m modestly exceeds $\sqrt{n}$—intuitively, this follows from the Birthday paradox, though making it formal requires appealing to uniformity testing. Therefore, the majority learner has a vanishing certifiable error rate starting around $m = \sqrt{n}$ samples. The OIG and ERM learners, in contrast, suffer constant error until their number of samples exceeds the number of minority labels, which is only slightly sublinear in n. To see why this is the case, note that any set of points smaller than the number of minority labels is shattered by the hypothesis class. Taking the disjoint union of this construction over all integers n, this essentially rules out any subquadratic bound on the blowup in sample complexity needed for ERM/OIG to catch up to the certifiable error rate of the majority learner.

We now present our positive result, which shows that a quadratic blowup in sample complexity—independent of the marginal—suffices to compete with certifiable semi-supervised guarantees.

Theorem 3.2 *For every domain and hypothesis class, the OIG learner is relatively* $\left(e, e\frac{m^2}{\eta}\right)$*-smart in the distribution-free setting.*

Theorem 3.2 stands in contrast to the impossibility result of (Darnstädt et al., 2013, Theorem 2), as described in Section 1.[5] This impossibility of smart learning can therefore be attributed to the challenge of certifying distribution-dependent error rates from unlabeled data. Taking the contrapositive, a simple corollary of Theorem 3.2 is that sound certification—by way of a learner/certifier pair for each distribution—of near-optimal distribution-fixed error rates suffices for smart learning.

We formally prove Theorem 3.2 in Section 3.1, but the high-level idea is as follows. Consider a learner $\mathcal{A}$ specialized to a marginal $\mathcal{D}$, and let $\mathcal{C}$ be its certifier. When allowed m samples, the certifier cannot distinguish between $\mathcal{D}$ and the uniform distribution $\mathcal{D}'$ on S, where S consists of $M = cm^2$ i.i.d. samples from $\mathcal{D}$ for a sufficiently large constant c. Intuitively, this follows from the Birthday paradox, though it takes some technical work to make it precise.[6] Since we require soundness of our certifiers, any m-sample certifiable error for $\mathcal{D}$ can be no better than the best error attainable on $\mathcal{D}'$ with m samples. The OIG learner—being dataset-by-dataset optimal in the leave-one-out sense—can be shown competitive with this when given roughly M i.i.d. samples from $\mathcal{D}$, as those correspond to what is effectively a constant fraction of the support of $\mathcal{D}'$. This yields the result.

It is natural to wonder whether ERM—or, for that matter, any learner that is simpler and typically more tractable than OIG—is also relatively smart with some finite (perhaps even quadratic) blowup in sample complexity. We however leave this question wide open.

Open Question 3.3 *Is ERM relatively smart in the distribution-free setting? Failing that, what about other natural and tractable learners?*

The sample complexities of ERM and OIG are closely related in traditional PAC learning, so it is tempting to suspect something similar here. The challenge in proving such a statement, however, comes from the fact that fine-grained dataset-by-dataset comparisons between OIG and ERM, or quantities like the VC dimension, appear bound to falter. It is known that there are hypothesis classes where ERM drastically trails the OIG learner in leave-one-out error on some datasets—consider for example the behaviors of hamming weight at most one on a large dataset, and an ERM

5. While phrased for a family of marginals, their result implies the same impossibility in the distribution-free setting. This is because the benchmark in smart learning (unlike ours) does not depend on the distribution class as a whole.
6. We note that a similar argument is employed in (Gollakota et al., 2023b, Theorem 6.2).

learner which breaks ties against the all-zero behavior. Similarly, there are hypothesis classes and arbitrarily large unlabeled datasets S where the VC dimension of induced behaviors is $|S|-1$, yet the OIG learner achieves zero leave-one-out error. As an example of this, consider the *parity* class on S, which ensures that the labels on S sum to $0 \pmod 2$. It therefore appears that any comparison between the error rates of the two learners cannot be argued dataset-by-dataset using leave-one-out arguments or the VC dimension. This suggests that either new proof approaches are needed, or ERM may not be relatively smart after all.

3.1. Formal Proof of Theorem 3.2

Let $\epsilon(.,.)$ be any certifiable error rate. Fix a distribution $\mathcal{D}$ and a sample size $m \geq 3$. We prove that $\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}, \frac{e}{\eta}m^2) \leq e^{1-\eta/3e}\epsilon(\mathcal{D},m) + \eta$ for every $\eta \in (0,1)$. Since $\epsilon(.,.)$ is certifiable, by definition there exists a learner $\mathcal{A} = \mathcal{A}_{\mathcal{D}}$ catered to $\mathcal{D}$ and a corresponding sound certifier $\mathcal{C} = \mathcal{C}_{\mathcal{A}}$ such that (i) $\epsilon_{\mathcal{A}}(\mathcal{D},m) \leq \mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{C}(S)] \leq \epsilon(\mathcal{D},m)$ and (ii) for any distribution $\mathcal{D}'$, $\epsilon_{\mathcal{A}}(\mathcal{D}',m) \leq \mathbb{E}_{S\sim\mathcal{D}'^m}[\mathcal{C}(S)]$, where the second fact crucially relies on soundness of the certifier.

It is easy to observe that the process of sampling m i.i.d. samples from $\mathcal{D}$ is equivalent to sampling a larger multi-set $S \sim \mathcal{D}^M$ of size M and then drawing a multi-set $T \sim_{\mathsf{nr}} \mathcal{D}_S^m$ of m samples from S uniformly at random *without replacement*. Denote by $\Gamma := \frac{M(M-1)\ldots(M-m+1)}{M^m}$ the probability of observing no duplicates when making m independent draws (with replacement) from a uniform distribution on M points. Let $M = cm^2$ for a constant $c = c(\eta)$ to be chosen later. Then,

$$\begin{aligned}\epsilon(\mathcal{D},m) \geq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathcal{C}(S)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}\mathop{\mathbb{E}}_{T\sim_{\mathsf{nr}}\mathcal{D}_S^m}[\mathcal{C}(T)] &\geq \frac{1}{\Gamma}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}\mathop{\mathbb{E}}_{T\sim\mathcal{D}_S^m}[\mathcal{C}(T)] + 1 - \frac{1}{\Gamma} \\ &\geq \frac{1}{\Gamma}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}[\epsilon_{\mathcal{A}}(\mathcal{D}_S,m)] + 1 - \frac{1}{\Gamma},\end{aligned} \tag{1}$$

where the first inequality is due to property (i), the second inequality is by simple algebraic manipulation and the fact that certifier outputs are at most 1, and the last inequality is due to property (ii).

For any distribution $\mathcal{D}'$ and sample size m' denote by $\epsilon^*(\mathcal{D}',m') = \inf_{\mathcal{A}}\epsilon_{\mathcal{A}}(\mathcal{D}',m')$ the minimum error of any learner on $\mathcal{D}'$ when the input is m' i.i.d. draws from $\mathcal{D}'$ (a.k.a. the optimal distribution-fixed error for $\mathcal{D}'$ with m' samples). Building on Equation (1) we get

$$\begin{aligned}\epsilon(\mathcal{D},m) \geq \frac{1}{\Gamma}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}[\epsilon_{\mathcal{A}}(\mathcal{D}_S,m)] + 1 - \frac{1}{\Gamma} &\geq \frac{1}{\Gamma}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}[\epsilon^*(\mathcal{D}_S,m)] + 1 - \frac{1}{\Gamma} \\ &\geq \frac{1}{\Gamma}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^M}[\epsilon^*(\mathcal{D}_S,M-1)] + 1 - \frac{1}{\Gamma},\end{aligned} \tag{2}$$

where the last inequality follows from the monotonicity of optimal distribution-fixed rates.

Next we upperbound the error of the OIG learner on $\mathcal{D}$ in terms of the expected optimal distribution-fixed error on an empirical distribution $\mathcal{D}_S$, for a sample S drawn i.i.d. from $\mathcal{D}$. This is articulated in the following Lemma.

Lemma 3.4 *For any distribution $\mathcal{D}$ and sample size m, denote by $\epsilon^*(\mathcal{D},m)$ the minimum error of any learner on $\mathcal{D}$ when given m i.i.d. samples. Then $\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D},M-1) \leq e\,\mathbb{E}_{S\sim\mathcal{D}^M}[\epsilon^*(\mathcal{D}_S,M-1)]$.*

To prove Lemma 3.4 we first use the reduction from transductive to PAC learning employed in the proof of (Asilis et al., 2025, Lemma A.1) (similar also to (Asilis et al., 2024, Lemma 34)) to conclude that for each unlabeled dataset S of size M there is a learner with worst-case transductive error on S bounded by $e \cdot \epsilon^*(\mathcal{D}_S, M-1)$. We then invoke the fact that OIG minimizes worst-case

transductive error (recall Definition A.3) to conclude that the transductive error of OIG on S is also at most $e \cdot \epsilon^*(\mathcal{D}_S, M-1)$. The lemma then follows from the standard leave-one-argument which upperbounds any learner's expected error in the PAC setting with $M-1$ training samples by its expected transductive error on datasets of size M drawn from the same distribution. Remaining technical details for proof of Lemma 3.4 appear in Appendix D.

Combining Lemma 3.4 with Equation (2) we get that

$$\epsilon(\mathcal{D}, m) \geq \frac{1}{e \cdot \Gamma} \epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}, M-1) + 1 - \frac{1}{\Gamma}.$$

Finally, for $c > 2$ and $m \geq 3$ we use the fact that $1 - 1/c \leq \Gamma \leq e^{-1/3c}$ to conclude that

$$\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}, M-1) \leq e \cdot \Gamma \cdot \epsilon(\mathcal{D}, m) + e - e \cdot \Gamma \leq e^{1-1/3c} \epsilon(\mathcal{D}, m) + e/c.$$

Setting $c = \frac{e}{\eta}$ then yields $\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}, M-1) \leq e^{1-\eta/3e} \epsilon(\mathcal{D}, m) + \eta$, as desired. ■

4. Distribution-Free Setting: An Impossibility for all Learners

Section 3 established that in order to compete with every certifiable semi-supervised guarantee, a quadratic blowup in sample complexity is necessary for the OIG and ERM learners, and sufficient for OIG. Can a different learner do better? Theorem 3.1 leaves open this possibility, as the "hard marginals" from that construction are all amenable to a single learner, namely majority. Nonetheless, we show by way of a more intricate hypothesis class that the answer in general is no, and a quadratic blowup in sample complexity is in fact the best possible. We prove the following theorem.

Theorem 4.1 *There exists a hypothesis class on a countable domain such that for every pair of constants α and β and every function $\sigma(m, \eta) = O_\eta(m^{2-\beta})$, no learner is relatively (α, σ)-smart.*

The formal proof of Theorem 4.1 can be found in Section 4.1, but the high-level idea is as follows. For each integer n and parameter β, we build a set system $\mathcal{S} = \mathcal{S}(n, \beta)$ on a large domain with $|S| = n$ for each $S \in \mathcal{S}$ and $|S \cap S'| \leq n^{1-O(\beta)} = o(n)$ for distinct $S, S' \in \mathcal{S}$. A hypothesis class $\mathcal{H} = \mathcal{H}(n, \beta) = \{h_S : S \in \mathcal{S}\}$ is then defined so that h_S is effectively a random behavior on S, but constant elsewhere (and therefore constant on most of any other $S' \in \mathcal{S}$).

First, we show that a learner catered to the uniform distribution on some known $S \in \mathcal{S}$ achieves vanishing error with very few samples m, and can certify this guarantee using uniformity testing on S when m modestly exceeds $\sqrt{|S|} = \sqrt{n}$. (The certifier outputs the trivial bound, effectively eschewing any guarantee, if the uniformity tester fails). This is possible because only h_S is "interesting" on S, whereas all other $h_{S'}$ with $S' \neq S$ are constant on all but $|S \cap S'| = o(n)$ points in S, and can therefore be "spotted" with few samples (much akin to our analyses of the majority learner from Theorem 3.1). Second, we show that $\mathcal{H}$ can be made sufficiently rich to shatter every set of size roughly $n^{1-O(\beta)}$, which prohibits any meaningful learning with less than that many samples when S is unknown. Taken together, these two properties imply the theorem for a fixed parameter β and number of samples $m \approx \sqrt{n}$. The theorem then follows by taking the disjoint union of the hypothesis classes $\mathcal{H}(n, \beta)$ for all integers n and a countable sequence of βs tending to 0.

4.1. Formal Proof of Theorem 4.1

Fix an arbitrary constant $\beta \in (0, \frac{1}{8})$. We will construct a set system $\mathcal{S}$ and a hypothesis class $\mathbb{H}$ that witness the impossibility of relative smartness of any learner for $\sigma(m, \eta) = O_\eta(m^{2-14\beta})$.

Using the probabilistic method, we construct set systems $\mathcal{S}(n) = \mathcal{S}(n, \beta)$ and hypothesis classes $\mathcal{H}(n) = \mathcal{H}(n, \beta)$ satisfying certain properties for sufficiently large $n \in \mathbb{N}$. We then take the disjoint union over them to create $\mathcal{S}$ and $\mathbb{H}$.

The following lemma, proved in Appendix E.1, summarizes the properties of $\mathcal{S}(n)$ we exploit.

Lemma 4.2 *There exists an absolute constant $C = C(\beta) \in \mathbb{N}$ such that for all $n \geq C$, there exists a set system $\mathcal{S}(n)$ on a universe $\mathcal{U}(n)$ with the following properties.*

(i) $|\mathcal{U}(n)| = \lceil n^{1+\beta} \rceil$ *and* $|\mathcal{S}(n)| = \lceil \exp\left(\frac{1}{4} n^{1-\beta/2}\right) \rceil$

(ii) $|S| = n$ *for all* $S \in \mathcal{S}(n)$

(iii) $|S \cap S'| \leq n^{1-\beta/2}$ *for any distinct* $S, S' \in \mathcal{S}(n)$

(iv) Every subset $T \subset \mathcal{U}(n)$ *with size* $|T| \leq 2n^{1-\beta}$ *is contained in at least* $\frac{1}{2} \exp\left(\frac{1}{8} n^{1-\beta/2}\right)$ *many sets in* $\mathcal{S}(n)$*, i.e.,* $|\{S \in \mathcal{S}(n) : T \subseteq S\}| \geq \frac{1}{2} \exp\left(\frac{1}{8} n^{1-\beta/2}\right)$.

For each $T \subseteq \mathcal{U}(n)$ we use $\mathcal{S}_T := \{S \in \mathcal{S}(n) : T \subseteq S\}$ to denote the collection of sets in $\mathcal{S}(n)$ containing T. Our hypothesis class $\mathcal{H}(n)$ satisfies properties outlined in the following lemma, whose proof can be found in Appendix E.2.

Lemma 4.3 *There exists an absolute constant $C' = C'(\beta) \in \mathbb{N}, C' \geq C$ such that for all $n \geq C'$ there exists a hypothesis class $\mathcal{H}(n) := \{h_S : S \in \mathcal{S}(n)\} \subseteq \{0,1\}^{\mathcal{U}(n)}$ with the following properties.*

(i) For all $S \in \mathcal{S}(n)$ and $x \notin S$, $h_S(x) = 1$.

(ii) For every subset $T \subset \mathcal{U}(n)$ with $|T| \leq 2n^{1-\beta}$ and every binary labeling $b \in \{0,1\}^{|T|}$,

$$(1 - n^{-\beta}) \frac{|\mathcal{S}_T|}{2^{|T|}} \leq |\{h_S \in \mathcal{H}(n) : S \in \mathcal{S}_T, h_S|T = b\}| \leq (1 + n^{-\beta}) \frac{|\mathcal{S}_T|}{2^{|T|}}.$$

The hypothesis class $\mathbb{H} = \mathbb{H}_\beta$ is now defined by taking the disjoint union of $\mathcal{H}(n)$ over all $n \geq C'$ as follows. Define the universe $\mathcal{U} := \biguplus_{n \geq C'} \mathcal{U}(n)$ and let $\mathcal{S} := \biguplus_{n \geq C'} \mathcal{S}(n)$. Let $\overline{\mathcal{H}}(n)$ be the extension of $\mathcal{H}(n)$ where each $h \in \overline{\mathcal{H}}(n)$ is constant 1 on $\mathcal{U} \setminus \mathcal{U}(n)$. Finally, let $\mathbb{H} := \biguplus_{n \geq C'} \overline{\mathcal{H}}(n)$.

Define $m(n) = \lfloor n^{1/2+3\beta} \rfloor$ and $\xi(n) = n^{-\beta/2}$. Observe that for each $S \in \mathcal{S}(n)$, each $h_{S'}$ with $S' \neq S$ assigns the label 1 to every point in S except possibly to $S \cap S'$, which under $\mathcal{D}_S$ has probability mass at most $|S \cap S'|/n \leq n^{-\beta/2} = \xi(n)$. Therefore, one of h_S or the constant 1 predictor incurs error at most $\xi(n)$ under $\mathcal{D}_S$, and validating between them using $m \geq \Omega\left(\frac{\ln(1/\xi(n))}{\xi(n)}\right)$ samples suffices for guaranteeing error $O(\xi(n))$ when $\mathcal{D}_S$ is fixed and known.

Sound certification of such a bound faces two interrelated obstacles, however: (a) The certifier must recognize when the true distribution $\mathcal{D}'$ is far from $\mathcal{D}_S$ and output a valid upperbound on the learner's error, and (b) the learner's error on each $\mathcal{D}'$ must be sufficiently bounded away from 1 to accommodate the difficulty inherent to (a). We therefore construct, for each $S \in \mathcal{S}(n)$, a learner $\mathcal{A}_S$ (Algorithm 1) and certifier $\mathcal{C}_S$ (Algorithm 2) that are jointly designed to satisfy both requirements. For (b), we ensure errors are bounded away from 1 for *every* distribution by adding the constant 0 and constant 1 predictors into consideration by the learner. In other words, $\mathcal{A}_S$ validates between h_S and the majority learner. For (a), much like in our proof of Theorem 3.1, $\mathcal{C}_S$ employs uniformity

Algorithm 1. Learner $\mathcal{A}_S(T)(x)$ trained on sample T given test point x:

1. Randomly split T into two equal parts T_1 and T_2, i.e., $T = T_1 \uplus T_2$ with $|T_1| = \left\lceil \frac{|T|}{2} \right\rceil$.
2. If $L(h_S, T_2) \leq L(\mathcal{A}_{\mathsf{Maj}}(T_1), T_2)$ then return $h_S(x)$;
3. Else return $\mathcal{A}_{\mathsf{Maj}}(T_1)(x)$.

Algorithm 2. Certifier $\mathcal{C}_S(T)$ run on (unlabeled) input sample T:

1. Let $n = |S|$ and $O = \mathsf{MTestUnif}_S(\xi(n), \xi(n), T)$. ▷ Algorithm 4 for testing against $\mathcal{D}_S$
2. If $|T| \geq m(n)$ and $O = 1$ then return $6\xi(n)$;
3. Else return 1.

testing to recognize distributions close to $\mathcal{D}_S$, and outputs small values only in those cases. Since uniformity testers are only accurate for sufficiently large m, our soundness guarantee is restricted to $m \geq m(n)$. In summary, $\mathcal{C}_S$ soundly certifies error $O(\xi(n))$ for $\mathcal{A}_S$ when given $m \geq m(n)$ samples from $\mathcal{D}_S$. This is articulated in the following lemma, formal proof of which can be found in Appendix E.3.

Lemma 4.4 *There exists an absolute constant $C'' = C''(\beta) \geq C'$ such that for all $n \geq C''$ the rate*

$$\epsilon_n(\mathcal{D}, m) = \begin{cases} 7n^{-\beta/2} & \mathcal{D} \in \{\mathcal{D}_S : S \in \mathcal{S}(n)\}, m \geq m(n) \\ 1 & \textit{otherwise.} \end{cases}$$

is certifiable in the distribution-free setting.

It remains to show that no learner can be smart relative to the above certifiable error rate. We prove that any learner $\mathcal{A}$, on average over all distributions in $\{\mathcal{D}_S : S \in \mathcal{S}(n)\}$, has error at least $1/2 - 2\xi(n)^2$ when given at most $n^{1-\beta}$ samples. This implies that for each $m \leq n^{1-\beta}$, there exists a set $S^* = S^*_{n,m} \in \mathcal{S}(n)$ such that $\epsilon_{\mathcal{A}}(\mathcal{D}_{S^*}, m) \geq 1/2 - 2\xi(n)^2$. The key idea is that for any sample of size at most $n^{1-\beta}$ and any fixed test point, roughly half of the distributions $\mathcal{D}_S$ consistent with this sample–test pair—namely, those for which S contains both the sample and the test point, and h_S is consistent with the sample—label the test point differently from the learner's prediction. Unlike the lower bound of PAC learning where the distributions are defined on shattered sets and the desired conclusion follows directly from the definition, the sets in our system can be much larger than the size of shattered sets. Property (ii) in Lemma 4.3 is introduced precisely to ensure that such a guarantee continues to hold in this setting. This is formally captured in the following lemma, the proof of which is relegated to Appendix E.4.

Lemma 4.5 *Let $\mathcal{A} : (\mathcal{U} \times \{0,1\})^* \times \mathcal{U} \to \{0,1\}$ be any learner. For any $n \geq C''$ and $m \leq n^{1-\beta}$ there exists a set $S^* = S^*_{n,m} \in \mathcal{S}(n)$ such that $\epsilon_{\mathcal{A}}(\mathcal{D}_{S^*}, m) \geq 1/2 - 2n^{-\beta}$.*

Lemmas 4.4 and 4.5 conclude that any fully-supervised learner requires at least $n^{1-\beta}$ many samples for its error to approximate—within any multiplicative constant α and any additive constant

$\eta < \frac{1}{2}$, both independent of n—the (certifiable) error rates of $\mathcal{D}_S$-fixed learners for $S \in \mathcal{S}(n)$. It follows that no learner is relatively $(\alpha, \sigma(m, \eta))$-smart for any constant α and $\sigma(m) = O_\eta(m^{2-14\beta})$. We conclude by taking a disjoint union of countably many $\mathbb{H}_\beta$—with disjoint domains $\mathcal{X}_\beta$—for a sequence of β tending to zero (e.g. $\beta \in \{1/i : i \in \mathbb{N}, i > 8\}$), where each $h \in \mathbb{H}_\beta$ is canonically extended to $\mathcal{X}$ by setting $h(x) = 1$ for $x \notin \mathcal{X}_\beta$. See Remark C.3 for more detail. ■

5. Distribution Families

In this section we examine relatively smart learning in distribution-family settings. We begin with the simple observation that our proof of Theorem 3.2 only required a simple closure property of the distribution family—namely, closure under taking empirical distributions. This captures, for example, any distribution family defined by a family of allowable manifolds, where the only restriction is that each distribution's support must lie inside one of the manifolds.

Corollary 5.1 *Fix an arbitrary hypothesis class on some domain. Let $\mathbb{D}$ be a distribution family with the following closure property: For every $\mathcal{D} \in \mathbb{D}$ and every finite set S contained in the support of $\mathcal{D}$, we also have $\mathcal{D}_S \in \mathbb{D}$. The OIG learner is relatively $\left(e, e\frac{m^2}{\eta}\right)$-smart over $\mathbb{D}$.*

Beyond such "simple" distribution families, relatively smart learning can start to behave quite differently from the distribution-free setting. There are distribution families where smart (and therefore relatively smart) learning is possible, but neither OIG nor ERM is relatively smart. Furthermore, there are families where relatively smart learning is not possible at all. The following pair of theorems summarize these findings.

Theorem 5.2 *There is a hypothesis class on a countable domain, and a countable distribution family on that domain, such that the following hold with respect to the family: There is a $(1, m)$-smart learner, yet neither OIG nor ERM is relatively smart.*[7]

Theorem 5.3 *There is a hypothesis class on a countable domain, and a countable distribution family $\mathbb{D}$ on that domain, such that no learner is relatively smart over $\mathbb{D}$.*

Theorem 5.2 is proved via a straightforward reduction from smart learning: We take a non-smartly-learnable family and augment the unlabeled data with a "tag" which is required to uniquely identify the marginal. This yields a family which is smartly learnable, with certifiable rates. Any algorithm which ignores the tags, such as ERM or OIG, is relatively smart for the "tagged" problem only if it is smart for the original problem. We relegate the straightforward proof to Appendix F.1. The tag construction should be viewed as a deliberately simple example of a broader phenomenon. Whenever marginal distributions in the family can be learned well from unlabeled data—of which recognizing a unique tag is the most trivial example—and moreover this aids in subsequent supervised learning, algorithms such as ERM or OIG which do not perform such unsupervised learning of the marginal need not be relatively smart.

Theorem 5.3 is also by reduction from smart learning, but is somewhat more interesting so we include it here.

Proof of Theorem 5.3. To rule out relatively smart learning, we show that it suffices for $\mathbb{D}$ to satisfy two properties: (i) $\mathbb{D}$ does not admit a smart learner, and (ii) $\mathbb{D}$ is *well separated* in that there

7. We mean that there exist no α and σ such that the learner is relatively (α, σ)-smart. (Recall Definition 2.5).

exists an absolute constant $c \in [0,1)$ so that $\mathbb{P}_{\mathcal{D}'}[\mathsf{supp}(\mathcal{D})] < c$ for any distinct pair of distributions $\mathcal{D}, \mathcal{D}' \in \mathbb{D}$.

Consider $\mathbb{D}$ satisfying (i) and (ii). For $\mathcal{D} \in \mathbb{D}$ let $\mathcal{A} = \mathcal{A}_{\mathcal{D}}$ be its optimal distribution-fixed learner, and let $\mathcal{B} = \mathcal{B}_{\mathcal{D}}$ be the learner which on m samples guesses randomly with probability $2c^m$ and otherwise invokes $\mathcal{A}$. Clearly, $\mathcal{B}$ correctly guesses any label with probability at least $2c^m/2 = c^m$, so $\epsilon_{\mathcal{B}}(\mathcal{D}', m) \leq 1 - \frac{2c^m}{2} = 1 - c^m$ for any distribution $\mathcal{D}'$. Consider the following certifier $\mathcal{C} = \mathcal{C}_{\mathcal{B}}$: Given a sample S of size m, if $S \subseteq \mathsf{supp}(\mathcal{D})$ then output $\epsilon_{\mathcal{B}}(\mathcal{D}, m) = (1 - 2c^m)\epsilon_{\mathcal{A}}(\mathcal{D}, m) + c^m$, otherwise output 1. This is sound for $\mathcal{B}$: If $S \sim \mathcal{D}^m$ then $\mathbb{E}[\mathcal{C}(S)] = \epsilon_{\mathcal{B}}(\mathcal{D}, m)$, and if $S \sim \mathcal{D}'^m$ for $\mathcal{D}' \in \mathbb{D}$ not equal to $\mathcal{D}$ then $\mathbb{E}[\mathcal{C}(S)] > 1 - c^m \geq \epsilon_{\mathcal{B}}(\mathcal{D}', m)$ by property (ii). This witnesses a certifiable error rate of $\epsilon_{\mathcal{B}}(\mathcal{D}, m) \leq \epsilon_{\mathcal{A}}(\mathcal{D}, m) + c^m$ for $\mathcal{D}$, which tends to the optimal distribution-fixed error $\epsilon_{\mathcal{A}}(\mathcal{D}, m)$ for large m. Since there is no smart learner over $\mathbb{D}$ by property (i), whereas certifiable errors tend to distribution-fixed errors additively as m grows large, there can be no relatively smart learner over $\mathbb{D}$.

It remains to exhibit a countable family on a countable domain satisfying (i) and (ii). The family of uniform distributions $\bigcup_{n,\beta}\{\mathcal{D}_S : S \in \mathcal{S}(n,\beta)\}$ from Theorem 4.1 is such a family. Indeed, (i) is because for $S \in \mathcal{S}(n,\beta)$ there is a distribution-fixed learner for $\mathcal{D}_S$ with error on the order of $n^{-O(\beta)}$ with $n^{O(\beta)}$ samples,[8] whereas absent knowledge of S no nontrivial learning is possible until the number of samples exceeds $n^{1-O(\beta)}$ (Lemma 4.5). As for (ii), it follows by construction since sets in $\mathcal{S}(n,\beta)$ have vanishingly small pairwise intersections (Lemma 4.2). ■

Finally, we reflect on how the difficulty of relatively smart learning varies as a function of the distribution family. Growing the family can certainly make relatively smart learning harder, as it always does for PAC learning, smart learning, and most other learning paradigms. Somewhat unusually, however, we show that the opposite can also occur for relatively smart learning. This non-monotonicity phenomenon is summarized in the following corollary of Theorems 5.3 and 3.2.

Corollary 5.4 *There is a hypothesis class on a countable domain, and three distribution families $\mathbb{D}_1 \subset \mathbb{D}_2 \subset \mathbb{D}_3$, such that $\mathbb{D}_1$ and $\mathbb{D}_3$ admit a relatively smart learner, but $\mathbb{D}_2$ does not.*

Proof Let $\mathbb{D}_2$ be the distribution family from Theorem 5.3, and fix its associated domain and hypothesis class. Let $\mathbb{D}_3$ consist of all distributions on the domain, and let $\mathbb{D}_1 = \{\mathcal{D}\}$ for any single distribution $\mathcal{D} \in \mathbb{D}_2$. It is clear that any singleton family such as $\mathbb{D}_1$ is smartly learnable, whereas Theorem 3.2 yields a relatively-smart learner for $\mathbb{D}_3$. ■

Corollary 5.4 might appear paradoxical until we recall that our benchmark in relatively smart learning—unlike most other learning paradigms—depends on the distribution family as a whole. While our desired learner may be burdened by having to handle more distributions, so too are the certifier/learner pairs with which it must compete. Specifically, for any learner $\mathcal{A}$ catered to a marginal $\mathcal{D}$, if we grow the distribution family then a certifier $\mathcal{C}$ for $\mathcal{A}$ must now be sound with respect to a larger class of marginal distributions, pushing the certifiable error on $\mathcal{D}$ upwards!

8. The learner which validates between h_S and the constant 1 predictor has error $\widetilde{O}\left(n^{-O(\beta)} + \frac{1}{m}\right)$ on m samples.

References

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems*, 28, 2015.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Regularization and Optimal Multiclass Learning. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pages 260–310. PMLR, June 2024. URL https://proceedings.mlr.press/v247/asilis24a.html. ISSN: 2640-3498.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Proper learnability and the role of unlabeled data. In *Algorithmic Learning Theory*, pages 112–133. PMLR, 2025.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pages 33–44, 2008.

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, September 1991. ISSN 0304-3975. doi: 10.1016/0304-3975(91)90026-X. URL https://www.sciencedirect.com/science/article/pii/030439759190026X.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-03358-9. OCLC: ocm64898359.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 287–316. PMLR, 2014.

Malte Darnstädt and Hans Ulrich Simon. Smart pac-learners. *Theoretical Computer Science*, 412 (19):1756–1766, 2011.

Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Unlabeled data does provably help. In *30th International Symposium on Theoretical Aspects of Computer Science (STACS 2013)*, pages 185–196. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. SIAM, 2014.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, pages 41–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2018.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, and Sankeerth Rao. Communication and memory efficient testing of discrete distributions. In *Conference on Learning Theory*, pages 1070–1106. PMLR, 2019a.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chic. J. Theor. Comput. Sci*, 25:1–21, 2019b.

Ilias Diakonikolas, Daniel Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 36:39470–39490, 2023.

R.M. Dudley, S.R. Kulkarni, T. Richardson, and O. Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Transactions on Information Theory*, 40(3):883–885, May 1994. ISSN 1557-9654. doi: 10.1109/18.335898. URL https://ieeexplore.ieee.org/document/335898. Conference Name: IEEE Transactions on Information Theory.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*, 11(19):625–660, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/erhan10a.html.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Amir Globerson, Roi Livni, and Shai Shalev-Shwartz. Effective Semisupervised Learning on Manifolds. In *Proceedings of the 2017 Conference on Learning Theory*, pages 978–1003. PMLR, June 2017. URL https://proceedings.mlr.press/v65/globerson17a.html. ISSN: 2640-3498.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity*, 7(20), 2000.

A Gollakota, AR Klivans, K Stavropoulos, and A Vasilyan. An efficient tester-learner for halfspaces. 12th International Conference on Learning Representations (ICLR 2024), 2024.

Aravind Gollakota, Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Tester-learners for halfspaces: Universal algorithms. *Advances in Neural Information Processing Systems*, 36: 10145–10169, 2023a.

Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1657–1670, 2023b.

Alexander Golovnev, David Pal, and Balazs Szorenyi. The information-theoretic value of unlabeled data in semi-supervised learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2328–2336. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/golovnev19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1500–1518. PMLR, June 2019. URL https://proceedings.mlr.press/v99/gopfert19a.html. ISSN: 2640-3498.

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-Functions on Randomly Drawn Points. *Information and Computation*, 115(2):248–292, December 1994. ISSN 0890-5401. doi: 10.1006/inco.1994.1097. URL https://www.sciencedirect.com/science/article/pii/S0890540184710972.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2887–2943. PMLR, 2024.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Rattana Pukdee, Dylan Sam, J Zico Kolter, Maria-Florina F Balcan, and Pradeep Ravikumar. Learning with explanation constraints. *Advances in neural information processing systems*, 36:49883–49926, 2023.

Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1643–1656, 2023.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014. ISBN 978-1-139-95274-3.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782, 1557-7317. doi: 10.1145/1968.1972. URL https://dl.acm.org/doi/10.1145/1968.1972.

Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Appendix A. Definitions: ERM and OIG

We use the standard definition of empirical risk minimization.

Definition A.1 *Empirical Risk Minimization (ERM) over a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ selects as its predictor an arbitrary $h \in \mathcal{H}$ minimizing average error on the training data. Since we are in the realizable setting, this is an arbitrary hypothesis consistent with all the training examples. As usual, when evaluating the error of ERM we assume worst-case tie-breaking among consistent hypotheses.*

The One-inclusion Graph (OIG) learner is best described in the context of transductive learning, as originally employed by Haussler et al. (1994). We also find the exposition in (Daniely and Shalev-Shwartz, 2014; Asilis et al., 2024) helpful.

Definition A.2 (Transductive Learning) *In the transductive model of learning for a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, an adversary selects a collection of unlabeled instances $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and a labeling hypothesis $h \in \mathcal{H}$. Let $S_h = ((x_1, h(x_1)), \ldots, (x_n, h(x_n))$ be the labeled sequence. Then, a data point x_i is selected uniformly at random from S as the test point, and the remaining data points and their labels, denoted by $S_h^{(-i)} = \{(x_j, y_j)\}_{j \neq i}$, are revealed to the learner $\mathcal{A}$. In other words, the learner is trained on $S_h^{(-i)}$ and tested on the label of x_i.*

The transductive error rate of a learner $\mathcal{A}$ is defined as

$$\epsilon_{\mathcal{A}}^{\mathsf{Trans}}(S_h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\mathcal{A}(S_h^{(-i)})(x_i) \neq h(x_i)].$$

Whenever the hypothesis class $\mathcal{H}$ is clear from context, we will overload notation and denote the worst-case transductive error of $\mathcal{A}$ on the unlabeled set S over any $h \in \mathcal{H}$ as

$$\epsilon_{\mathcal{A}}^{\mathsf{Trans}}(S) = \max_{h \in \mathcal{H}} \epsilon_{\mathcal{A}}^{\mathsf{Trans}}(S_h).$$

Note that any learner in the transductive model is well defined when realizable data is drawn i.i.d. (as in the PAC model) by taking S to be the union of the training and test points, then predicting accordingly. A leave-one-out argument upper bounds expected error over the distribution—where expectation is over i.i.d. draws of training and test data—by the worst-case transductive error.

Definition A.3 (OIG Learner) *The One-Inclusion Graph (OIG) learner for a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is that which minimizes worst-case transductive error $\epsilon_{\mathcal{A}}^{\mathsf{Trans}}(S)$, simultaneously for every unlabeled dataset $S \in \mathcal{X}^*$. Operationally, this learner can be described as follows for a given unlabeled dataset $S = (x_1, \ldots, x_n)$: Let G be the subgraph of the hypercube $\{0,1\}^n$ induced by the behaviors $\mathcal{H}|S$ of $\mathcal{H}$ on S, and find the (randomized) orientation of the edges of G which minimizes the maximum (expected) outdegree. When only a single behavior in $\mathcal{H}|S$ is consistent with the training data $\{(x_j, y_j)\}_{j \neq i}$, the algorithm predicts y_i accordingly. When both labels are possible for y_i given the training labels, then the oriented edge between the two corresponding behaviors determines the label chosen by the algorithm; in particular, the label y_i corresponding to the behavior pointed to by the directed edge is chosen.*

Remark A.4 *Note that we allow the OIG algorithm to randomize over orientations. This can be equivalently described as a fractional orientation minimizing the maximum fractional out-degree, which is then subjected to independent randomized rounding. The distinction between integral (i.e. deterministic) and fractional (i.e. randomized) orientations is not particularly consequential, as the associated integrality gap—and therefore the multiplicative gap in error—is merely a factor of 2. However, the latter does have the distinction of obtaining the optimal transductive error rate. These subtleties are discussed in Asilis et al. (2024).*

Appendix B. Uniformity Testing Preliminaries

In this section, we review the background of uniformity testing. There is a substantial body of work on uniformity testing, with many algorithms developed for uniformity and identity testing under a variety of modeling assumptions (Goldreich and Ron, 2000; Paninski, 2008; Diakonikolas et al., 2014; Acharya et al., 2015; Valiant and Valiant, 2017; Diakonikolas et al., 2018, 2019b,a; Acharya et al., 2020). In this work, we adopt the collision-based approach of Goldreich and Ron (2000), which we refer to as the standard uniformity tester.

We first recall the definition of total variation distance for countably-supported distributions.

Definition B.1 (Total Variation distance) *Let $\mathcal{D}$ and $\mathcal{Q}$ be two probability distributions defined over a countable domain $\mathcal{X}$. The TV distance between $\mathcal{D}$ and $\mathcal{Q}$ is defined as*

$$\mathsf{d}_{\mathsf{TV}}(\mathcal{D},\mathcal{Q}) = \frac{1}{2}\|\mathcal{D}-\mathcal{Q}\|_1 = \frac{1}{2}\sum_{x\in\mathcal{X}}|\mathcal{D}[x]-\mathcal{Q}[x]|.$$

We now describe the standard collision-based uniformity tester of Goldreich and Ron (2000), denoted by $\mathsf{TestUnif}$. This tester assumes that the unknown distribution $\mathcal{D}$ is supported on the same finite domain as the reference uniform distribution. This assumption about the support is crucial for relating collision statistics to the distance from uniformity. The formal description of $\mathsf{TestUnif}$ appears in Algorithm 3 and its performance guarantee is stated below.

Lemma B.2 (Diakonikolas et al. (2019b)) *Denote $m_{\mathsf{Test}}(n,\xi,\delta) := \frac{18\cdot 64\sqrt{n}\ln(2/\delta)}{\xi^2}$. For any set Y of size $|Y|=n$, $\xi,\delta\in(0,1)$, any distribution $\mathcal{D}$ such that $\mathsf{supp}(\mathcal{D})\subseteq Y$, and any sample size $m \geq m_{\mathsf{Test}}(n,\xi,\delta)$ Algorithm 3 satisfies that*

- *If $\mathcal{D}=\mathcal{D}_Y$, $\mathsf{TestUnif}_Y(\xi,\delta,S)$ returns 1 with probability at least $1-\delta$ over $S\sim\mathcal{D}^m$*
- *If $\mathsf{d}_{\mathsf{TV}}(\mathcal{D},\mathcal{D}_Y)>\xi$, $\mathsf{TestUnif}_Y(\xi,\delta,S)$ returns 0 with probability at least $1-\delta$ over $S\sim\mathcal{D}^m$*

In our setting, we will consider distributions that may place nonzero mass outside the reference domain Y. To handle this, we rely on a simple but useful observation showing how total variation distance behaves under conditioning.

Lemma B.3 *Let $\mathcal{X}$ be a countable set and let $Y\subseteq\mathcal{X}$ be finite. Let $\mathcal{D},\mathcal{D}'$ be any pair of distributions supported on $\mathcal{X}$. We have that*

$$\mathsf{d}_{\mathsf{TV}}(\mathcal{D}_{|Y},\mathcal{D}') \geq \mathsf{d}_{\mathsf{TV}}(\mathcal{D},\mathcal{D}') - \mathcal{D}[\mathcal{X}\setminus Y].$$

Algorithm 3. Standard Uniformity Tester $\mathsf{TestUnif}_Y(\xi, \delta, S)$ for a set Y with input parameters $\xi, \delta \in (0,1)$ and sample S (Goldreich and Ron, 2000).

1. Let $n = |Y|$, $\ell = 18\ln(2/\delta)$, and $m' = |S|/\ell$.
2. Divide S into ℓ consecutive sub-samples $S_i = (x_{i1}, \ldots, x_{im'}), 1 \le i \le \ell$
3. Let $\text{TR} = \frac{1+2\xi^2}{n}$.
4. For $1 \le i \le \ell$ do:
   (a) Let $Z_i = \frac{1}{\binom{m'}{2}} |\{(j,k) : j < k, x_{ij} = x_{ik}\}|$.
   (b) If $Z_i < \text{TR}$ then let $\text{ACC}_i = 1$, else let $\text{ACC}_i = 0$.
5. If $\sum_i \text{ACC}_i \ge \ell/2$ then return 1, else return 0.

Proof This follows by the triangle inequality and the definition of total variation distance.

$$\mathsf{d_{TV}}(\mathcal{D}', \mathcal{D}) \le \mathsf{d_{TV}}(\mathcal{D}', \mathcal{D}_{|Y}) + \mathsf{d_{TV}}(\mathcal{D}_{|Y}, \mathcal{D}) = \mathsf{d_{TV}}(\mathcal{D}', \mathcal{D}_{|Y}) + \mathcal{D}[\mathcal{X} \setminus Y]$$

■

Lemma B.3 shows that if a distribution $\mathcal{D}$ places only a small amount of probability mass outside Y, then conditioning $\mathcal{D}$ on Y preserves total variation distance up to a small additive difference. Consequently, testing $\mathcal{D}$ for uniformity over Y can be reduced to testing the conditional distribution $\mathcal{D}_{|Y}$, provided that samples falling outside Y are explicitly detected. This observation motivates a simple wrapper around the standard uniformity tester, which rejects whenever a sample lies outside Y and otherwise invokes TestUnif (with modified parameters) on the sample. The modified uniformity tester MTestUnif appears in Algorithm 4, and its guarantees are summarized in the following lemma.

Lemma B.4 *Denote* $m_{\mathsf{Test}}(n, \xi, \delta) := \frac{18 \cdot 64\sqrt{n}\ln(2/\delta)}{\xi^2}$. *Let* $\mathcal{X}$ *be an arbitrary countable set. For any set* $Y \subseteq \mathcal{X}$ *of size* $|Y| = n$, *any* $\xi, \delta \in (0,1)$, *any distribution* $\mathcal{D}$ *on* $\mathcal{X}$, *and any sample size* $m \ge m_{\mathsf{Test}}(n, \xi/2, \delta)$, *the tester* $\mathsf{MTestUnif}_Y(\xi, \delta, S)$ *(Algorithm 4) satisfies the following:*

- *If* $\mathcal{D} = \mathcal{D}_Y$, $\mathsf{MTestUnif}_Y(\xi, \delta, S)$ *returns* 1 *with probability* $\ge 1 - \delta$ *over* $S \sim \mathcal{D}^m$.
- *If* $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_Y) > \xi$, $\mathsf{MTestUnif}_Y(\xi, \delta, S)$ *returns* 0 *with probability* $\ge 1 - \delta$ *over* $S \sim \mathcal{D}^m$.

Proof Let $m \ge m_{\mathsf{Test}}(n, \xi/2, \delta) = \frac{18 \cdot 64\sqrt{n}\ln(2/\delta)}{(\xi/2)^2}$. If $\mathcal{D} = \mathcal{D}_Y$ then observe that $S \subseteq Y$ and the first claim follows from the first guarantee of Lemma B.2. Moreover, if $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_Y) > \xi$ and $\mathcal{D}[\mathcal{X} \setminus Y] \le \xi/2$, we have from Lemma B.3 that $\mathsf{d_{TV}}(\mathcal{D}_{|Y}, \mathcal{D}_Y) \ge \mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_Y) - \mathcal{D}[\mathcal{X} \setminus Y] \ge \xi/2$. In this case if $S \not\subseteq Y$, then the tester returns 0, correctly rejecting $\mathcal{D}$. If $S \subseteq Y$, then we know that S is distributed i.i.d. according to $\mathcal{D}_{|Y}$. By our choice of m we have with probability $1 - \delta$ over $S \sim \mathcal{D}^m$ that the tester returns 0 correctly rejecting. Finally, if $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_Y) > \xi$ and $\mathcal{D}[\mathcal{X} \setminus Y] > \xi/2$ then since $m \ge \frac{2\ln(1/\delta)}{\xi}$ it is easy to check that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$ we have $S \not\subseteq Y$ in which case the tester returns 0. This completes the proof. ■

Algorithm 4. Modified uniformity tester $\mathsf{MTestUnif}_Y(\xi, \delta, S)$ for set Y with parameters $(\xi, \delta) \in (0,1)$ and input sample S:

1. If $S \subseteq Y$ then return $\mathsf{TestUnif}_Y(\xi/2, \delta, S)$; ▷ Algorithm 3 for testing against $\mathcal{D}_Y$
2. Else return 0.

Appendix C. Formal Proof of Theorem 3.1

Fix an arbitrary constant $\beta \in \left(0, \frac{1}{8}\right)$. We will construct a hypothesis class $\mathbb{H}$ for which OIG and ERM are not relatively smart for any $\sigma(m, \eta) = O_\eta(m^{2-14\beta})$. Consider the domain $\mathcal{X} \subseteq \mathbb{N} \times \mathbb{N}$ with $\mathcal{X} := \bigcup_{n \in \mathbb{N}} \mathcal{X}_n$ and $\mathcal{X}_n := \{n\} \times [n]$ for all $n \in \mathbb{N}$. We will sometimes refer to $\mathcal{X}_n$ as the *nth row* of the domain. Denote by $\mathcal{D}(n)$ the uniform distribution on the nth row $\mathcal{X}_n$. We will define a hypothesis class $\mathcal{H}(n)$ such that the probability under $\mathcal{D}(n)$ of minority label of any $h \in \mathcal{H}(n)$ is about $n^{-\beta}$. Define the functions $M, m : \mathbb{N} \to \mathbb{N}$ and $\xi :\to (0,1)$ as

$$M(n) = \lceil n^{1-\beta} \rceil, m(n) = \lfloor n^{\frac{1}{2}+3\beta} \rfloor, \text{ and } \xi(n) = M(n)/n \approx n^{-\beta}. \tag{3}$$

For $b \in \{0,1\}$ and $n \in \mathbb{N}$, let

$$\mathcal{H}^{(b)}(n) = \{h \in \{0,1\}^{\mathcal{X}} : \forall x \in \mathcal{X} \setminus \mathcal{X}_n, h(x) = 1 \text{ and } |\{x \in \mathcal{X}_n : h(x) = b\}| = M(n)\},$$

and define $\mathcal{H}(n) = \mathcal{H}^{(0)}(n) \cup \mathcal{H}^{(1)}(n)$. In other words, for each $h \in \mathcal{H}(n)$ there are exactly $M(n)$ points in row n with the minority label, whereas h is one everywhere outside that row. Note that $\xi(n)$ is the probability under $\mathcal{D}(n)$ of a minority label for any $h \in \mathcal{H}(n)$, which approaches 0 as n grows. We let $\mathbb{H} := \bigcup_{n \in \mathbb{N}} \mathcal{H}(n)$.

Let $\mathcal{A}_{\mathsf{Maj}}$ be the learner which ignores unlabeled data, and always predicts the most frequent label seen in training; formally, $\mathcal{A}_{\mathsf{Maj}}(S)(x') = \mathbb{1}\left[|\{(x,y) \in S : y = 1\}| \geq |S|/2\right]$. We show that $\mathcal{A}_{\mathsf{Maj}}$ has error $O(\xi(n))$ on $\mathbb{H}$ with respect to $\mathcal{D}(n)$ and sample sizes $m \geq m(n)$ and, more importantly, this error rate is certifiable. We prove the following.

Lemma C.1 *There exists an absolute constant $C = C(\beta) \in \mathbb{N}$ such that for all $n \geq C$, the rate*

$$\epsilon_n(\mathcal{D}, m) = \begin{cases} 4\xi(n) & \text{if } \mathcal{D} = \mathcal{D}(n) \text{ and } m \geq m(n) \\ 1 & \text{otherwise.} \end{cases}$$

is certifiable for $\mathcal{A}_{\mathsf{Maj}}$ in the distribution-free setting.

We prove Lemma C.1 by certifying the error rate of $\mathcal{A}_{\mathsf{Maj}}$. Specifically, for each row $n \in \mathbb{N}$ we exhibit a certifier $\mathcal{C}_{\mathsf{Maj},n}$ (Algorithm 5) satisfying $\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}, m) \leq \mathbb{E}_{S \sim \mathcal{D}^m}[\mathcal{C}_{\mathsf{Maj},n}(S)] \leq \epsilon_n(\mathcal{D}, m)$ for all $m \in \mathbb{N}$ and all distributions $\mathcal{D}$ on $\mathcal{X}$. Note that for distributions on the row of interest n, $\mathcal{A}_{\mathsf{Maj}}$ will only do well for distributions close to uniform, so our certifier must output small values exclusively for those. We therefore exploit uniformity testing—recapped in Appendix B—to distinguish the uniform distribution on row n from distributions a substantial total variation distance away. Our uniformity tester adds a simple wrapper around the standard uniformity tester $\mathsf{TestUnif}$ (Algorithm 3), which allows detecting distributions with support outside the desired row. The resulting modified uniformity tester $\mathsf{MTestUnif}$ is shown in Algorithm 4 with guarantees summarized

Algorithm 5. Certifier $\mathcal{C}_{\mathsf{Maj},n}(S)$ run on input sample S:

1. Let $O = \mathsf{MTestUnif}_{\mathcal{X}_n}(\xi(n), \xi(n), S)$. ▷ Algorithm 4 for testing against $\mathcal{D}(n)$
2. If $|S| \geq m(n)$ and $O = 1$ then return $3\xi(n)$, else return 1.

in Lemma B.4. The remaining technical details for proof of Lemma C.1 are fairly standard, and therefore relegated to Appendix C.1.

Having upper-bounded the certifiable error rate of the majority learner for $\mathcal{D}(n)$ with $m \geq m(n)$ samples (Lemma C.1), it remains to lower bound the error of ERM and OIG learners in the same regime. The proof of the following Lemma is fairly elementary and relegated to Appendix C.2.

Lemma C.2 *For the hypothesis class $\mathbb{H}$ we have $\epsilon_{\mathcal{A}_{\mathsf{ERM}}}(\mathcal{D}(n), m) \geq 1-2\xi(n)$ and $\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}(n), m) \geq 1/2 - 2n^{-\beta/2} = 1/2 - o(1)$ for all $m < M(n)$.*

For the distribution $\mathcal{D}(n)$, Lemmas C.1 and C.2 imply that it would take at least $M(n) \approx n^{1-\beta}$ samples for the error of OIG or ERM to approximate—to within any multiplicative constant α and any additive constant $\eta < \frac{1}{2}$, both independent of n—the certifiable error rate of the majority learner with $m(n) \approx n^{\frac{1}{2}+3\beta}$ samples. Since $M(n) = \omega\left((m(n))^{2-14\beta}\right)$ this rules out relative (α, σ)-smartness of OIG and ERM for any constant α and any $\sigma(m, \eta) = O_\eta(m^{2-14\beta})$. ■

Remark C.3 *The preceding proof of Theorem 3.1 may appear to require a different hypothesis class $\mathbb{H}_\beta$ for every choice of the parameter β. This can be avoided by taking the disjoint union of countably many $\mathbb{H}_\beta$—over disjoint countable domains $\mathcal{X}_\beta$—for a sequence of β values tending to zero. In more detail, for each $\beta \in B = \{1/i : i \in \mathbb{N}, i > 8\}$ we define $\mathbb{H}_\beta$ over its own distinct domain $\mathcal{X}_\beta$ as in the preceding proof, let $\mathcal{X} = \biguplus_{\beta \in B} \mathcal{X}_\beta$, and extend each $h \in \mathbb{H}_\beta$ to $\mathcal{X}$ canonically by setting $h(x) = 1$ for $x \notin \mathcal{X}_\beta$. The resulting single hypothesis class $\mathbb{H} = \biguplus_{\beta \in B} \mathbb{H}_\beta$ over the countable domain $\mathcal{X}$ serves to witness the impossibility result of Theorem 3.1, as needed.*

C.1. Proof of Lemma C.1

We will use the following lemma as upper bound on the error of the majority learner and lower bound on the error of ERM for learning from $\mathcal{H}(n)$ under the uniform distribution $\mathcal{D}(n)$.

Lemma C.4 *Let $\xi(n)$ be defined as in Equation (3). Assume n is sufficiently large such that $n > \frac{\log(2/\xi(n))}{2\xi(n)(1/2-\xi(n))^2}$. Then, for the hypothesis class $\mathbb{H}$ and sample sizes $\frac{\log(2/\xi(n))}{2(1/2-\xi(n))^2} \leq m \leq M(n)$ we have (i) $\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}(n), m) \leq 2\xi(n)$ and (ii) $\epsilon_{\mathcal{A}_{\mathsf{ERM}}}(\mathcal{D}(n), m) \geq 1 - 2\xi(n)$.*

Proof Fix any $h \in \mathbb{H}$. Let $b \in \{0, 1\}$ be the minority label of h on $\mathcal{X}_n$, i.e.,$|\{x \in \mathcal{X}_n : h(x) = b\}| \leq M(n)$ and thus either $h \notin \mathcal{H}(n)$ or $h \in \mathcal{H}^{(b)}(n)$. Denote $\mathcal{X}_b := \{x \in \mathcal{X}_n : h(x) = b\}$ and $\mathcal{X}_{1-b} = \mathcal{X}_n \setminus \mathcal{X}_b$. We know $\mathcal{D}(n)[\mathcal{X}_b] \leq M(n)/n = \xi(n)$ and $\mathcal{D}[\mathcal{X}_{1-b}] \geq 1 - \xi(n)$. Let $\delta \in (0, 1)$ and S be any sample of size $m \geq \frac{\ln(2/\delta)}{2(1/2-\xi(n))^2}$. In the event that $|S \cap \mathcal{X}_{1-b}| > |S|/2$, we know $\mathcal{A}_{\mathsf{Maj}}$ outputs $1 - b$ on every input and clearly has error $\mathcal{D}(n)[\mathcal{X}_b] \leq \xi(n)$. We bound the event that the

majority label in sample is not $1-b$ by Hoeffding's inequality

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}(n)^m}\left[|S\cap\mathcal{X}_{1-b}|\le\frac{|S|}{2}\right] &= \mathop{\mathbb{P}}_{S\sim\mathcal{D}(n)^m}\left[\frac{1}{m}|S\cap\mathcal{X}_{1-b}|\le\frac{1}{2}\right]\\ &\le \mathop{\mathbb{P}}_{S\sim\mathcal{D}(n)^m}\left[\left|(1-\xi(n))-\frac{1}{m}|S\cap\mathcal{X}_{1-b}|\right|\ge\frac{1}{2}-\xi(n)\right]\\ &\le 2\exp(-2m(1/2-\xi(n))^2)\le\delta.\end{aligned}$$

Therefore we can conclude that $\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D},m)\le\xi(n)+\delta$. On the other hand if $m\le M(n)$, there always exists an ERM hypothesis that will incorrectly pick b as the majority label. This is due the fact that $|S|\le M(n)$ and there always exists a hypothesis $\widehat{h}$ consistent with S that has $|\{x\in\mathcal{X}_n:\widehat{h}(x)=1-b\}|=M(n)$. It is clear that such a hypothesis has error at least $1-2\xi(n)$. Letting $\delta=\xi(n)$ concludes the result. We also want to make sure such a value of m exists. The lower bound requirement on n is used to make sure that $\frac{\ln(2/\xi(n))}{2(1/2-\xi(n))^2}\le m<\xi(n)\cdot n=M(n)$. ■

We now restate Lemma C.1 and prove it.

Lemma C.5 (Restatement of Lemma C.1) *There exists an absolute constant $C=C(\beta)\in\mathbb{N}$ such that for all $n\ge C$, the rate*

$$\epsilon_n(\mathcal{D},m)=\begin{cases}4\xi(n) & \text{if } \mathcal{D}=\mathcal{D}(n) \text{ and } m\ge m(n)\\ 1 & \text{otherwise.}\end{cases}$$

is certifiable for $\mathcal{A}_{\mathsf{Maj}}$ in the distribution-free setting.

Proof Recall the definition of functions $M,m:\mathbb{N}\times\mathbb{N}$ and $\xi:\mathbb{N}\times(0,1)$

$$M(n)=\lceil n^{1-\beta}\rceil, m(n)=\lfloor n^{\frac{1}{2}+3\beta}\rfloor, \text{ and } \xi(n)=M(n)/n\approx n^{-\beta}.$$

We prove that $\mathcal{A}_{\mathsf{Maj}}$ certifiably achieves the error rate $\epsilon_n(.,.)$ with $\mathcal{C}_{\mathsf{Maj},n}$ being the sound certifier. To do so, we have to make sure that

$$\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D},m)\le\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathcal{C}_{\mathsf{Maj},n}(S)]\le\epsilon_n(\mathcal{D},m)$$

for all $\mathcal{D},m$. It is obvious that whenever $m<m(n)$, regardless of the distribution, the certifier always outputs 1, which is an upper bound on the error rate and also is equal to the certifiable error rate. We will have to consider three cases for the distribution $\mathcal{D}$ when $m\ge m(n)$. Throughout, we will be using the following immediate corollary of Lemma B.4 and always assume n is larger than the constant in the following to handle different cases.

Lemma C.6 *There exists an absolute constant $C_1=C_1(\beta)\in\mathbb{N}$ such that for any $n\ge C_1(\beta)$, we have for any $m\ge m(n)$,*

- *If $\mathcal{D}=\mathcal{D}(n)$, $\mathsf{MTestUnif}_{\mathcal{X}_n}(\xi(n),\xi(n),S)$ returns 1 with probability at least $1-\xi(n)$ over $S\sim\mathcal{D}^m$.*
- *If $\mathsf{d}_{\mathsf{TV}}(\mathcal{D},\mathcal{D}(n))>\xi(n)$, $\mathsf{MTestUnif}_{\mathcal{X}_n}(\xi(n),\xi(n),S)$ returns 0 with probability at least $1-\xi(n)$ over $S\sim\mathcal{D}^m$.*

Proof We can verify that for sufficiently large n, for $m(n) = \lfloor n^{\frac{1}{2}+3\beta} \rfloor$ as defined in Equation (3) in Section C we have

$$\lfloor n^{\frac{1}{2}+3\beta} \rfloor \geq \frac{18 \cdot 64\sqrt{n} \ln\left(2n^{\beta}\right)}{n^{-2\beta}} = \frac{18 \cdot 64\sqrt{n} \ln\left(\frac{2}{\xi(n)}\right)}{\xi^2} = m_{\mathsf{Test}}(n, \xi(n), \xi(n)). \tag{4}$$

This implies that any $m \geq m(n)$ satisfies the requirement of Lemma B.4 for $S = \mathcal{X}_n$ and $\xi = \xi(n)$, which concludes the proof. ■

We now discuss each case separately.

1. The distribution is $\mathcal{D}(n)$. We can verify that for sufficiently large n (and thus small $\xi(n)$), we have the following lower bounds on n and $m(n)$ in terms of $\xi(n)$

$$n \geq \frac{\ln\left(2n^{\beta}\right)}{2n^{-\beta}(\frac{1}{2} - n^{-\beta})^2} = \frac{\ln\left(\frac{2}{\xi(n)}\right)}{2\xi(n)(\frac{1}{2} - \xi(n))^2}, \tag{5}$$

and

$$m \geq m(n) \geq \frac{\ln\left(2n^{\beta}\right)}{2(\frac{1}{2} - n^{-\beta})^2} = \frac{\ln\left(\frac{2}{\xi(n)}\right)}{2(\frac{1}{2} - \xi(n))^2}. \tag{6}$$

Therefore, we can invoke Lemma C.4 to conclude that $\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}(n), m) \leq 2\xi(n) = 2M(n)/n$. Moreover, from Lemma C.6, we know that with probability at least $1-\xi(n)$ over $S \sim \mathcal{D}(n)^m$, $\mathsf{MTestUnif}_{\mathcal{X}_n}(\xi(n), \xi(n), S)$ (Algorithm 4) returns 1, accepting that the underlying distribution is $\mathcal{D}(n)$. Hence, with probability at least $1 - \xi(n)$ the certifier outputs $3\xi(n)$. Moreover, observe that $\mathcal{C}_{\mathsf{Maj},n}(S) \geq 3\xi(n)$ for any sample S and we have

$$\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}(n), m) \leq 3\xi(n) \leq \mathop{\mathbb{E}}_{S \sim \mathcal{D}(n)^m}[\mathcal{C}_{\mathsf{Maj},n}(S)] \leq 3\xi(n) + \xi(n) = 4\xi(n) = \epsilon_n(\mathcal{D}(n), m).$$

2. $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}(n)) \leq \xi(n)$. We will prove that the error of majority learner is always bounded by $3\xi(n)$ for n sufficiently large such that $\xi(n) < 1/4$. Let b denote the majority label of the labeling function $h^* \in \mathbb{H}$ under the distribution $\mathcal{D}$ and denote $p := \mathcal{D}[\{x : h^*(x) = b\}]$. Considering $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}(n)) \leq \xi(n)$, we observe that for $b = 1$, if $h^* \in \mathcal{H}(n)$, we have $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 1\}] = 1 - \xi(n)$ and $p \geq 1 - 2\xi(n)$. Note that $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 1\}] = 1 - \xi(n)$ comes from the fact that if by the sake of contradiction, we assume for h^* we have $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 1\}] = \xi(n)$, then we would get $p < 2\xi(n) < 1/2$, which is a contradiction to $b = 1$ being the majority label. If $h^* \notin \mathcal{H}(n)$, we have $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 1\}] = 1$ and $p \geq 1 - \xi(n)$. On the other hand, if $b = 0$, we have $h^* \in \mathcal{H}(n)$. To see this, assume by the sake of contradiction that $h^* \notin \mathcal{H}(n)$. Then we get that $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 0\}] = 0$ and thus $\mathcal{D}[\{x \in \mathcal{X}_n : h^*(x) = 0\}] \leq \xi(n) < 1/2$ which is a contradiction to 0 being the majority label. Therefore, $\mathcal{D}(n)[\{x \in \mathcal{X}_n : h^*(x) = 0\}] = 1 - \xi(n)$ and $p \geq 1 - 2\xi(n)$. In any case, we get that $1 - 2\xi(n) \leq p \leq 1$. Observe that $\mathbb{E}_{S \sim \mathcal{D}^m}[|\{(x, y) \in S : y = b\}|] \geq (1 - 2\xi(n)) \cdot m$ and similar to proof of Lemma C.4 we get from Chernoff's inequality that

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}[\text{Majority of labels in } S \text{ is } 1 - b] \leq 2\exp\left(-2m(1/2 - 2\xi(n))^2\right) \leq \xi(n),$$

where the last line follows from Equation (6) that holds for large n. Therefore, we get that

$$\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}, m) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbb{1}[\mathcal{A}_{\mathsf{Maj}}(x) \neq h^*(x)]] \leq \xi(n)p + (1-p) \leq \xi(n) + 2\xi(n) = 3\xi(n).$$

Taking into account that $\mathcal{C}_{\mathsf{Maj},n}(S) \geq 3\xi(n)$ for any S, we have $\mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{C}_{\mathsf{Maj},n}(S)] \geq 3\xi(n) \geq \epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}, m)$.

3. $\mathsf{d}_{\mathsf{TV}}(\mathcal{D}, \mathcal{D}(n)) > \xi(n)$**.** In this case, we can invoke Lemma C.6 to get that with probability at least $1-\xi(n)$ over $S \sim \mathcal{D}^m$, $\mathsf{MTestUnif}_{\mathcal{X}_n}(\xi(n), \xi(n), S)$ (Algorithm 4) rejects and outputs 0. Therefore, $\mathcal{C}_{\mathsf{Maj},n}$ outputs 1 with probability at least $1-\xi(n)$ and $\mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{C}_{\mathsf{Maj},n}(S)] \geq 1-\xi(n)$.

   We now find an upper bound on the error of $\mathcal{A}_{\mathsf{Maj}}$ on distribution $\mathcal{D}$. Let b denote the majority label of labeling function $h^* \in \mathbb{H}$ under distribution $\mathcal{D}$ and denote $p := \mathcal{D}[\{x : h^*(x) = b\}] \geq 1/2$. We get from Chernoff's inequality that

$$q := \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\text{Majority of labels in } S \text{ is } 1-b] \leq 2\exp\left(-2m(p-1/2)^2\right).$$

   We can therefore write that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m_{h^*}, x\sim\mathcal{D}}[\mathbb{1}[\mathcal{A}_{\mathsf{Maj}}(x) \neq h^*(x)]] = pq + (1-p)(1-q).$$

   Observe that since $p \geq 1/2$, we get from simple calculations that $pq + (1-p)(1-q) \leq p$. Now if $1/2 \leq p \leq 1/2 + \xi(n)$ we have $pq + (1-p)(1-q) \leq 1/2 + \xi(n)$, which is at most $3/4$ for large n where $\xi(n) < 1/4$. Moreover, we can verify that for sufficiently large n,

$$m(n) \geq \frac{\ln(\frac{16}{1+2\xi(n)})}{2(\xi(n))^2}.$$

   This implies that if $p > 1/2 + \xi(n)$ then we have $m \geq m(n) \geq \frac{\ln(\frac{16}{2p})}{2(p-1/2)^2}$ and, thus, $q \leq \frac{p}{4}$, which in turn proves $pq + (1-p)(1-q) \leq \frac{p^2}{4} + 1 - p \leq \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$. Since the choice of $h^* \in \mathbb{H}$ was arbitrary, we can get that

$$\epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}, m) = \sup_{h^*\in\mathcal{H}} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m_{h^*}, x\sim\mathcal{D}}[\mathbb{1}[\mathcal{A}_{\mathsf{Maj}}(x) \neq h^*(x)]] \leq \frac{3}{4}.$$

   This concludes that $\mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{C}_{\mathsf{Maj},n}(S)] \geq 1-\xi(n) \geq 3/4 \geq \epsilon_{\mathcal{A}_{\mathsf{Maj}}}(\mathcal{D}, m)$.

■

C.2. Proof of Lemma C.2

We know from Lemma C.4 that for the uniform distribution $\mathcal{D}(n)$ and the hypothesis class $\mathcal{H}(n)$ there are ERM learners with error rate $\epsilon_{\mathcal{A}_{\mathsf{ERM}}}(\mathcal{D}(n), m) \geq 1 - 2\xi(n)$ for all $m(n) \leq m < M(n)$ because $m < M(n)$ and that the Equations (5) and (6) imply that requirements of the lemma are satisfied for $m \geq m(n)$.

We now consider the OIG learner. For a multiset $S \in \mathcal{X}^*$, denote by $\mathsf{dom}(S) \subseteq \mathcal{X}$ the set of distinct elements in S and by $S_{(1)} \subseteq \mathsf{dom}(S)$ the set of all elements that appear exactly once

in S. Observe that for any multiset $S \in \mathcal{X}_n^*$ with $|S| \leq M(n)$, the set $\mathsf{dom}(S)$ is shattered by $\mathcal{H}(n)$. From Fact C.7, we know that for any such set there exists a maximum-out-degree minimizing orientation of the graph that sets the out-degree of every vertex to at least $\lfloor |S_{(1)}|/2 \rfloor$. Moreover, from Lemma C.8 we know that for any $m \leq M(n)$ we have with probability at least $1 - n^{-\beta/2}$ over $S \sim \mathcal{D}(n)^m$ that $|S_{(1)}| \geq m(1 - 2n^{-\beta/2})$. This implies that for any $h \in \mathcal{H}(n)$ with probability at least $1 - n^{-\beta/2} = 1 - o(1)$ over $S \sim \mathcal{D}(n)^m$, the trandsuctive error is at least $1/2 - n^{-\beta/2} = 1/2 - o(1)$. Combining this with the leave-one-out argument, we conclude that $\epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}(n), m) \geq 1/2 - 2n^{-\beta/2} = 1/2 - o(1)$ for all $m < M(n)$. ■

Fact C.7 *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class and $S \in \mathcal{X}^n$ a multiset of size n such that $\mathsf{dom}(S)$ is shattered by $\mathcal{H}$, i.e., $\mathcal{H}|\mathsf{dom}(S) = \{0,1\}^{\mathsf{dom}(S)}$. There exists a maximum-out-degree minimizing orientation of the OIG on S that sets the out-degree of every vertex to at least $\lfloor |S_{(1)}|/2 \rfloor$.*

Proof Let $d := |S_{(1)}|$. It is easy to verify that the one-inclusion graph of $\mathcal{H}|S$ is a disjoint union of d-dimensional hypercubes. This is due to the fact for any $u \in \{0,1\}^n$, the edges connected to u correspond exactly to the instances that appear exactly once in S. Moreover, any node u may only connect to nodes that are consistent with u on $S \setminus S_{(1)}$. It is therefore, enough to show that for the d-dimensional hypercube, there exists a maximum-out-degree minimizing orientation that sets the out-degree of every vertex to $\lfloor d/2 \rfloor$. We now prove this fact.

First note that the total number of edges in the d-dimensional hypercube is $d \cdot 2^{d-1}$ and any orientation of the graph must have maximum degree at least $\lceil d/2 \rceil$. We now describe an orientation that achieves maximum out-degree at most $\lceil d/2 \rceil$. For any edge $e = (u, v)$ in graph connecting nodes $u, v \in \{0,1\}^d$, let $i_e \in [d]$ be the instance such that $u(i_e) \neq v(i_e)$. We orient the edge towards the node u if the parity of u equals i_e mod 2 and we orient it to v otherwise. Obviously, each node u will have out-degree at least $\lfloor d/2 \rfloor$ and at most $\lceil d/2 \rceil$. Finally, note that this is a valid orientation since $e = (u, v)$ already implies that u and v have different parities. ■

Lemma C.8 *Let $m \leq M(n)$. We have with probability at least $1 - n^{-\beta/2} = 1 - o(1)$ over $S \sim \mathcal{D}(n)^m$ that $|S_{(1)}| \geq m(1 - 2n^{-\beta/2}) = m(1 - o(1))$.*

Proof Let Z denote the set of pairs of instances in the multiset $S = (x_1, \ldots, x_m)$ that are duplicates, i.e., $Z = \{1 \leq i < j \leq n : x_i = x_j\}$. We can verify that $\mathbb{E}_{S \sim \mathcal{D}(n)^m}[|Z|] = \binom{m}{2}\frac{1}{n}$. We apply Markov's inequality to conclude that

$$\mathbb{P}\Big[Z \geq n^{-\beta/2} m\Big] \leq \frac{m^2}{2n} \cdot \frac{1}{mn^{-\beta/2}} \leq \frac{1}{2n^{\beta/2}},$$

where the last line follows from the fact that $m \leq M(n) = n^{1-\beta}$. Note that if we remove all the instances in Z from S, we get the set of all instances of S that only appear once. Therefore, $|S_{(1)}| \geq |S| - 2|Z|$ and we get that with probability at least $1 - n^{-\beta/2}$ we have $|S_{(1)}| \geq m(1 - 2n^{-\beta/2})$. This concludes the proof. ■

Appendix D. Missing Proofs From Section 3.1

D.1. Proof of Lemma 3.4

Fix a multiset $S = (x_1, \ldots, x_M)$ of M instances. Let $\mathcal{A}_S$ be the learner with $\epsilon_{\mathcal{A}_S}(\mathcal{D}_S, M-1) = \epsilon^*(\mathcal{D}_S, M-1)$. We will now prove that there exists another learner $\mathcal{A}'$ such that

$$\epsilon^{\mathsf{Trans}}_{\mathcal{A}'}(S) \leq e \cdot \epsilon_{\mathcal{A}_S}(\mathcal{D}_S, M-1) = e \cdot \epsilon^*(\mathcal{D}_S, M-1),$$

On any set $S_h^{(-i)}$ and test point x, the learner $\mathcal{A}'$ will draw a set T of $|S_h^{(-i)}| = M-1$ i.i.d. samples from the uniform distribution on $S_h^{(-i)}$, which we denote by $\mathcal{D}_{h,-i}$. It then predicts $\mathcal{A}'(x) = \mathcal{A}_S(T)(x)$. Asilis et al. (2025) prove that in the process of sampling $M-1$ instances from the uniform distribution $\mathcal{D}_S$, the probability that a point $x \in S$, e.g., x_i, is not sampled is at least $1/e$. For any $h \in \mathcal{H}$, denote by $\mathcal{D}_h$ the uniform distribution on S. This implies that

$$\begin{aligned}
\epsilon^{\mathsf{Trans}}_{\mathcal{A}'}(S) = \max_{h \in \mathcal{H}} \epsilon^{\mathsf{Trans}}_{\mathcal{A}'}(S_h) &= \max_{h \in \mathcal{H}} \frac{1}{M} \sum_{i=1}^{M} 1\{\mathcal{A}'(S_h^{(-i)})(x_i) \neq h(x_i)\} \\
&= \max_{h \in \mathcal{H}} \frac{1}{M} \sum_{i=1}^{M} \mathop{\mathbb{E}}_{T \sim \mathcal{D}_{h,-i}^{M-1}} [1\{\mathcal{A}_S(T)(x_i) \neq h(x_i)\}] \\
&\leq \max_{h \in \mathcal{H}} e \cdot \frac{1}{M} \sum_{i=1}^{M} \mathop{\mathbb{E}}_{T \sim \mathcal{D}_{h}^{M-1}} [1\{\mathcal{A}_S(T)(x_i) \neq h(x_i)\}] \\
&= e \cdot \max_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{T \sim \mathcal{D}_{h}^{M-1}} \left[\frac{1}{M} \sum_{i=1}^{M} 1\{\mathcal{A}_S(T)(x_i) \neq h(x_i)\} \right] \\
&= e \cdot \epsilon_{\mathcal{A}_S}(\mathcal{D}_S, M-1) \\
&= e \cdot \epsilon^*(\mathcal{D}_S, M-1).
\end{aligned}$$

We now rely on the fact that the one-inclusion-graph learner achieves the optimal transductive error on any sample S (see Definition A.3 and Remark A.4), and therefore $\epsilon^{\mathsf{Trans}}_{\mathcal{A}_{\mathsf{OIG}}}(S) \leq \epsilon^{\mathsf{Trans}}_{\mathcal{A}'}(S) \leq e \cdot \epsilon^*(\mathcal{D}_S, M-1)$. Moreover, observe that by a simple leave-one-out argument we have

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^M} \left[\epsilon^{\mathsf{Trans}}_{\mathcal{A}_{\mathsf{OIG}}}(S) \right] = \mathop{\mathbb{E}}_{S \sim \mathcal{D}^M} \left[\max_{h \in \mathcal{H}} \epsilon^{\mathsf{Trans}}_{\mathcal{A}_{\mathsf{OIG}}}(S_h) \right] \geq \max_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^M} \left[\epsilon^{\mathsf{Trans}}_{\mathcal{A}_{\mathsf{OIG}}}(S_h) \right] = \epsilon_{\mathcal{A}_{\mathsf{OIG}}}(\mathcal{D}, M-1).$$

■

Appendix E. Missing Proofs from Section 4

E.1. Proof of Lemma 4.2

We use a probabilistic method argument. Let $\mathcal{U}$ be a universe of size $|\mathcal{U}| = \lceil n^{1+\beta} \rceil$. Pick $k := \lceil \exp(\frac{1}{4} n^{1-\beta/2}) \rceil$ many subsets uniformly at random from all the subsets of $\mathcal{U}$ of size n. We calculate the expected size of the intersection of any such pair by writing

$$\mathbb{E}[|S_i \cap S_j|] = \sum_{u \in \mathcal{U}} \mathbb{E}[1\{u \in S_i \wedge u \in S_j\}] = \sum_{u \in \mathcal{U}} \left(\frac{n}{|\mathcal{U}|} \right)^2 = \frac{n^2}{|\mathcal{U}|} \in \left(\frac{5}{6} n^{1-\beta}, n^{1-\beta} \right),$$

where the last step follows from the fact that $n^{1+\beta} \le |\mathcal{U}| \le \frac{6}{5}n^{1+\beta}$ for $n \ge 7$. Since the random variables $1\{u \in S_i \wedge u \in S_j\}$ are negatively associated we can now apply a Chernoff and union bound to conclude that

$$\begin{aligned}
\mathbb{P}\Big[\forall 1 \le i < j \le k,\, |S_i \cap S_j| \le n^{1-\beta/2}\Big] &\ge 1 - \sum_{1\le i<j\le k} \mathbb{P}\Big[|S_i \cap S_j| > n^{1-\beta/2}\Big] \\
&\ge 1 - k^2\, \mathbb{P}\Big[|S_i \cap S_j| > n^{1-\beta} \cdot (1 + n^{\beta/2} - 1)\Big] \\
&\ge 1 - k^2\, \mathbb{P}\Big[|S_i \cap S_j| > \mathbb{E}[|S_i \cap S_j|] \cdot (1 + n^{\beta/2} - 1)\Big] \\
&\ge 1 - k^2 \exp\left(-\frac{(n^{\beta/2}-1)^2}{2 + n^{\beta/2} - 1} \cdot \frac{5}{6} \cdot n^{1-\beta}\right) \\
&\ge 1 - k^2 \exp\left(-\frac{5}{6} \cdot \frac{64}{90} \frac{(n^{\beta/2})^2}{n^{\beta/2}} n^{1-\beta}\right) \\
&\ge 1 - k^2 \exp\left(-\frac{16}{27} \cdot n^{1-\beta/2}\right)
\end{aligned}$$

where we used the fact that for large n, we have both $n^{\beta/2} - 1 \ge \frac{8}{9}n^{\beta/2}$ and $1 + n^{\beta/2} \le \frac{10}{9}n^{\beta/2}$. Now since we have $k^2 \le 2\exp(\frac{1}{2}n^{1-\beta/2})$, we get

$$\mathbb{P}\Big[\forall i, j \in [k],\, |S_i \cap S_j| \le n^{1-\beta/2}\Big] \ge 1 - k^2 \exp\left(-n^{1-\beta/2}\right) \ge 1 - 2\exp\left(-\frac{1}{12}n^{1-\beta/2}\right). \quad (7)$$

We then analyze the requirement needed for property (iv). Take any subset $T \subset \mathcal{U}$ of size $|T| \le 2n^{1-\beta}$. For any $i \in [k]$ we have that

$$\mathbb{P}[T \subset S_i] = \frac{\binom{|\mathcal{U}|-|T|}{|S_i|-|T|}}{\binom{|\mathcal{U}|}{|S_i|}} = \frac{|S_i|(|S_i|-1)\ldots(|S_i|-|T|+1)}{|\mathcal{U}|(|\mathcal{U}|-1)\ldots(|\mathcal{U}|-|T|+1)} \ge \left(\frac{|S_i|-|T|}{|\mathcal{U}|-|T|}\right)^{|T|}.$$

Therefore, noting that $|T| \le 2n^{1-\beta}$, we can write for large n that

$$\mathbb{P}[T \subset S_i] \ge \left(\frac{n - 2n^{1-\beta}}{n^{1+\beta}}\right)^{2n^{1-\beta}} = \left(\frac{n^\beta - 2}{n^{2\beta}}\right)^{2n^{1-\beta}} \ge \left(\frac{1}{2n^\beta}\right)^{2n^{1-\beta}} = \exp\left(-2n^{1-\beta}\ln(2n^\beta)\right).$$

We can then conclude by linearity of expectation that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in[k]} \mathbb{1}[T \subset S_i]\right] &\ge k \exp\left(-2n^{1-\beta}\ln(2n^\beta)\right) \\
&\ge \exp\left(\frac{1}{4}n^{1-\beta/2} - 2n^{1-\beta}\ln(2n^\beta)\right) \\
&\ge \exp\left(n^{1-\beta/2}\left(\frac{1}{4} - 2n^{-\frac{\beta}{2}}\ln(2n^\beta)\right)\right) \\
&\ge \exp\left(\frac{1}{8}n^{1-\beta/2}\right),
\end{aligned}$$

where the last inequality follows from the fact that for large n, we have that $2n^{-\frac{\beta}{2}}\ln(2n^{\beta}) < 1/8$. Note that from a Chernoff bound we have that

$$\mathbb{P}\left[\sum_{i\in[k]} \mathbb{1}[T\subset S_i] < \frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right] \leq \exp\left(-\frac{1}{8}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right)$$

We apply a union bound over all subsets of size at most $2n^{1-\beta}$ to conclude that

$$\begin{aligned}
&\mathbb{P}\left[\forall T\subset\mathcal{U} : \sum_{i\in[k]} \mathbb{1}\{T\subset S_i\} \geq \frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right]\\
&= 1-\mathbb{P}\left[\exists T\subset\mathcal{U} : \sum_{i\in[k]} \mathbb{1}\{T\subset S_i\} < \frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right]\\
&\geq 1-\sum_{T\subset\mathcal{U}:|T|\leq 2n^{1-\beta}} \mathbb{P}\left[\sum_{i\in[k]} \mathbb{1}\{T\subset S_i\} < \frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right]\\
&\geq 1-2n^{1-\beta}\binom{2n^{1+\beta}}{2n^{1-\beta}}\mathbb{P}\left[\sum_{i\in[k]} \mathbb{1}\{T\subset S_i\} < \frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right]\\
&\geq 1-2n^{1-\beta}\left(en^{2\beta}\right)^{2n^{1-\beta}}\exp\left(-\frac{1}{8}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right),
\end{aligned}$$

where we used the fact that $|\mathcal{U}| \leq 2n^{1+\beta}$ in the second inequality. We can continue writing

$$\begin{aligned}
&1-2n^{1-\beta}\left(en^{2\beta}\right)^{2n^{1-\beta}}\exp\left(-\frac{1}{8}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right)\\
&= 1-\exp\left(\ln(2)+(1-\beta)\ln(n)+(2+4\beta\ln(n))n^{1-\beta}\right)\exp\left(-\frac{1}{8}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right)\\
&\geq 1-\exp\left(2(2+4\beta\ln(n))n^{1-\beta}-\frac{1}{8}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right) \qquad (8)\\
&\geq 1-\exp\left(-\frac{1}{16}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)\right)\\
&\geq 1-\exp\left(-\frac{1}{2}n^{1-\beta/2}\right),
\end{aligned}$$

where the first and second inequalities hold since for sufficiently large n we have $\ln(2)+(1-\beta)\ln(n) \leq (2+4\beta\ln(n))n^{1-\beta}$, and $2(2+4\beta\ln(n))n^{1-\beta} \leq \frac{1}{16}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)$.

Taking a union bound over Equation (7) and (8) we can conclude that with probability at least $1-3\exp\left(-\frac{1}{12}n^{1-\beta/2}\right)$ the intersection of every pair is at most $n^{1-\beta/2}$ and also every subset of universe of size at most $2n^{1-\beta}$ is contained in at least $\frac{1}{2}\exp\left(\frac{1}{8}n^{1-\beta/2}\right)$ many other sets. This proves the existence of a set system with the claimed properties. ■

E.2. Proof of Lemma 4.3

We will use a probabilistic method to show such a hypothesis class exist. In particular, for every $S \in \mathcal{S}(n)$, pick a random hypothesis h_S that is constant 1 outside S and the labeling on S is chosen uniformly at random from all the $2^{|S|}$ possible labelings. Clearly this satisfies property (i). Recall that $|S| = n$ and $|\mathcal{U}(n)| = \lceil n^{1+\beta} \rceil$. We show the a randomly chosen labeling of h_S for all $S \in \mathcal{S}(n)$ satisfies the claimed properties with non-zero probability.

To prove property (ii) we first show that with high probability over the random labeling, all the hypotheses $h_S, S \in \mathcal{S}(n)$ are unique, that is, for any distinct pair of $S, S' \in \mathcal{S}(n)$ we have $h_S \neq h'_S$. Since the hypothesis h_S is constant 1 on any $S' \setminus S$, it is sufficient to prove that for all $S \in \mathcal{S}(n)$, we have $|\{x \in S : h_S(x) = 0\}| \geq n/3$ because $S \cap S' \leq n^{1-\beta/2}$. We know for any S that $\mathbb{E}[|\{x \in S : h_S(x) = 0\}|] = n/2$. Taking a Chernoff and union bound we get that

$$\begin{aligned}
\mathbb{P}\left[\forall S \in \mathcal{S}(n), |\{x \in S : h_S(x) = 0\}| \geq \frac{n}{3}\right] &\geq 1 - \sum_{S \in \mathcal{S}(n)} \mathbb{P}\left[|\{x \in S : h_S(x) = 0\}| < \frac{n}{3}\right] \\
&\geq 1 - |\mathcal{S}(n)| \exp\left(-\frac{n}{36}\right) \\
&\geq 1 - \exp\left(\frac{1}{2}n^{1-\beta/2} - \frac{n}{36}\right) \\
&\geq 1 - \exp\left(-\frac{1}{2}n^{1-\beta/2}\right),
\end{aligned} \tag{9}$$

where we used the fact that for large n, we have $n/36 \geq n^{1-\beta/2}$.

We turn to proving property (ii). For any $T \subset \mathcal{U}(n)$ with $|T| \leq 2n^{1-\beta}$, the set is contained in $\mathcal{S}_T$ many sets and the labeling of each set is picked uniformly at random. For any fixed labeling $b \in \{0,1\}^{|T|}$, let $Z_{T,b}$ denote the random variable $|\{S \in \mathcal{S}_T, h_S | T = b\}|$. We know from property (iv) in Lemma 4.2 that $\mathcal{S}_T \geq \lfloor \frac{1}{2} \exp\left(\frac{1}{8}n^{1-\beta/2}\right) \rfloor \geq \frac{1}{4} \exp\left(\frac{1}{8}n^{1-\beta/2}\right)$ and, therefore, we have $\mathbb{E}[Z_{T,b}] = \frac{|\mathcal{S}_T|}{2^{|T|}} \geq 2^{-2n^{1-\beta}} \frac{1}{4} \exp\left(\frac{1}{8}n^{1-\beta/2}\right)$. We apply a Chernoff bound to conclude that

$$\begin{aligned}
\mathbb{P}\left[\left|Z_{T,b} - \frac{|\mathcal{S}_T|}{2^{|T|}}\right| \geq n^{-\beta}\frac{|\mathcal{S}_T|}{2^{|T|}}\right] &\leq 2\exp\left(-\frac{|\mathcal{S}_T|}{3 \cdot 2^{|T|}} n^{-2\beta}\right) \\
&\leq 2\exp\left(-\frac{n^{-2\beta}}{24}\exp\left(\frac{1}{8}n^{1-\beta/2} - 2\ln(2)n^{1-\beta}\right)\right) \\
&= 2\exp\left(-\frac{1}{24}\exp\left(\frac{1}{8}n^{1-\beta/2} - 2\beta\ln(n) - 2\ln(2)n^{1-\beta}\right)\right) \\
&\leq 2\exp\left(-\frac{1}{24}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right),
\end{aligned}$$

where the last inequality is due to the fact that $2\beta\ln(n) + 2\ln(2)n^{1-\beta} \leq \frac{1}{16}n^{1-\beta/2}$ for sufficiently large n.

Taking a union bound over all labelings b and all sets T, we get that

$$\begin{aligned}
&\mathbb{P}\left[\exists T, b : \left|Z_{T,b} - \frac{|\mathcal{S}_T|}{2^{|T|}}\right| \geq n^{-\beta}\frac{|\mathcal{S}_T|}{2^{|T|}}\right] \leq \sum_{\substack{T\subset\mathcal{U}(n), b\in\{0,1\}^{|T|}:\\ |T|\leq 2n^{1-\beta}}} \mathbb{P}\left[\left|Z_{T,b} - \frac{|\mathcal{S}_T|}{2^{|T|}}\right| \geq n^{-\beta}\frac{|\mathcal{S}_T|}{2^{|T|}}\right] \\
&\leq 4n^{1-\beta}\binom{2n^{1+\beta}}{2n^{1-\beta}}2^{2n^{1-\beta}}\exp\left(-\frac{1}{24}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right) \\
&\leq 4n^{1-\beta}\left(en^{2\beta}\right)^{2n^{1-\beta}}2^{2n^{1-\beta}}\exp\left(-\frac{1}{24}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right) \\
&\leq \exp\left(\ln(4) + (1-\beta)\ln(n) + 2(1+\ln(2)+2\beta\ln(n))n^{1-\beta}\right)\exp\left(-\frac{1}{24}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right) \\
&\leq \exp\left(-\frac{1}{48}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right),
\end{aligned} \tag{10}$$

where in the last inequality we used the fact that for large sufficiently large n, we have

$$\ln(4) + (1-\beta)\ln(n) + 2(1+\ln(2)+2\beta\ln(n))n^{1-\beta} \leq \frac{1}{48}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)$$

Taking another union bound with Equation (9), we conclude that with probability at least $1 - \exp\left(-\frac{1}{48}\exp\left(\frac{1}{16}n^{1-\beta/2}\right)\right) - \exp\left(-\frac{1}{2}n^{1-\beta/2}\right)$, all h_S are unique and, therefore, $Z_{T,b} = |\{h_S \in \mathcal{H}(n) : S \in \mathcal{S}_T, h_S|T = b\}|$ for all T, b and property (ii) is satisfied for all subsets of size at most $2n^{1-\beta}$. Observe that for the large n regime we are considering, property (ii) also implies that every labeling of T induces many hypotheses and that T is shattered. This implies the existence of the claimed labeling of the set-system and finishes the proof. ■

E.3. Proof of Lemma 4.4

We will prove that the certifier is both sound and upper bounded by the certifiable error rate. Formally, we prove that

$$\epsilon_{\mathcal{A}_S}(\mathcal{D}, m) \leq \mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}[\mathcal{C}_S(T)] \leq \epsilon_n(\mathcal{D}, m).$$

The claim is obvious when $m < m(n)$ where we always have $\epsilon_{\mathcal{A}_S}(\mathcal{D}, m) \leq \mathbb{E}_{T\sim\mathcal{D}^m}[\mathcal{C}_S(T)] = \epsilon_n(\mathcal{D}, m) = 1$.

Let $h^* \in \mathbb{H}$ be the labeling function. We will use the following throughout the proof to bound the error of the learner $\mathcal{A}_S$. For large n and $m \geq m(n)$, using a Chernoff and union bound we have

$$\mathop{\mathbb{P}}_{T_2\sim\mathcal{D}_{h^*}^{m/2}}[\exists h \in \{h_S, \mathcal{A}_{\mathsf{Maj}}(T_1)\},\ |L(h, T_2) - L(h, \mathcal{D}_{h^*})| > \xi(n)] \leq 4\exp\left(-m\xi(n)^2\right) \leq \xi(n), \tag{11}$$

where the last inequality is due to the fact that for large m, we have $m \geq \frac{\ln(4/\xi(n))}{\xi(n)^2}$.

We now consider three cases based on the underlying distribution. We will use the following instantiation of Lemma B.4 which can be proven similar to Lemma C.6 since Equation (4) still holds in this setting. We will assume that n is larger than the constant in the following lemma for the rest of the proof.

Lemma E.1 *There exists an absolute constant $C_1 = C_1(\beta) \in \mathbb{N}$ such that for any $n \geq C_1(\beta)$, we have for any $S \in \mathcal{S}(n)$ and $m \geq m(n)$,*

- *If $\mathcal{D} = \mathcal{D}_S$, $\mathsf{MTestUnif}_S(\xi(n), \xi(n), T)$ returns 1 with probability at least $1 - \xi(n)$ over $T \sim \mathcal{D}^m$*
- *If $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_S) > \xi(n)$, $\mathsf{MTestUnif}_S(\xi(n), \xi(n), T)$ returns 0 with probability at least $1 - \xi(n)$ over $T \sim \mathcal{D}^m$*

Let $r := |\{(x, y) \in T_1 : y = 1\}|$ be the number of instances in T_1 with label 1. We now discuss each case for the distribution separately.

1. $\mathcal{D} = \mathcal{D}_S$ for some $S \in \mathcal{S}(n)$. From Lemma E.1 we know with probability at least $1 - \xi(n)$ over $T \sim \mathcal{D}^m$, $\mathsf{MTestUnif}_S(\xi(n), \xi(n), T)$ accepts the underlying distribution as $\mathcal{D}_S$ and thus $O = 1$. Therefore, we have $\mathbb{E}_{T \sim \mathcal{D}^m}[\mathcal{C}_S(T)] \leq \xi(n) + 6\xi(n) = 7\xi(n)$.

   We now bound $\epsilon_{\mathcal{A}_S}(\mathcal{D}, m)$ for $m \geq m(n)$. Two possibilities can happen, either (1) $h^* = h_S$ or (2) $h^* = h_{S'}$ for some $S' \in \mathcal{S}, S' \neq S$. In the first possibility, h_S will have zero error under $\mathcal{D}_S$ and obviously $L(h_S, T_2) = L(h_S, \mathcal{D}_{h^*}) = 0$.

   In the second possibility where $h^* = h_{S'}$ for some $S' \neq S$ we know from property (i) in Lemma 4.3 that $h_{S'}$ is 1 on S except on the intersection of S and S' which has size at most $n^{1-\beta/2}$ based on property (iii) in Lemma 4.2. Therefore, $\mathbb{E}_{T_1}[r] \geq \frac{|S| - |S \cap S'|}{|S|} \cdot |T_1| \geq (1 - n^{-\beta/2})|T_1|$. Therefore, from a Chernoff bound, for sufficiently large n, we get that

$$\begin{aligned}
\mathop{\mathbb{P}}_{T}\left[r \leq \frac{1}{2} \cdot |T_1|\right] &= \mathop{\mathbb{P}}_{T}\left[r \leq \frac{1}{2(1 - n^{-\beta/2})} \cdot (1 - n^{-\beta/2})|T_1|\right] \\
&\leq \mathop{\mathbb{P}}_{T}\left[r \leq (1 - \frac{1}{3}) \cdot (1 - n^{-\beta/2})|T_1|\right] && \left(\frac{1}{2(1 - n^{-\beta/2})} \leq \frac{2}{3}\right) \\
&\leq \exp\left(-\frac{1}{18}(1 - n^{-\beta/2})|T_1|\right) \leq \exp\left(-\frac{m}{18 \cdot 2 \cdot 4}\right) && \left((1 - n^{-\beta/2}) \geq \frac{1}{2}\right),
\end{aligned}$$

   where we used the fact that $|T_1| \geq \lfloor \frac{|T|}{2} \rfloor \geq \frac{m}{4}$. It is easy to verify in the large n and $m \geq m(n)$ regime we have $\exp\left(-\frac{m}{144}\right) \leq n^{-\beta/2} \leq \xi(n)$. Therefore, with probability at least $1 - \xi(n)$ we get that $r > \frac{1}{2} \cdot |T_1|$ and $\mathcal{A}_{\mathsf{Maj}}(T_1)(x) = 1$ for all $x \in \mathcal{U}(n)$ and thus $L(\mathcal{A}_{\mathsf{Maj}}(T_1), \mathcal{D}_{h^*}) \leq \frac{|S \cap S'|}{|S|} \leq n^{-\beta/2} \leq \xi(n)$.

   Let $\widehat{h} \in \arg\min_{h \in \{h_S, \mathcal{A}_{\mathsf{Maj}}(T_1)\}} L(h, \mathcal{D}_{h^*})$. We proved in above that $L(\widehat{h}, \mathcal{D}_{h^*}) \leq \xi(n)$ with probability at least $1 - \xi(n)$. Noting that $\mathcal{A}_S(T) \in \arg\min_{h \in \{h_S, \mathcal{A}_{\mathsf{Maj}}(T_1)\}} L(h, T_2)$ and taking a union bound of the above failure with Equation (11), we get that with probability at least $1 - 2\xi(n)$ over $T \sim \mathcal{D}^m$

$$L(\mathcal{A}_S(T), \mathcal{D}_{h^*}) \leq L(\mathcal{A}_S(T), T_2) + \xi(n) \leq L(\widehat{h}, T_2) + \xi(n) \leq L(\widehat{h}, \mathcal{D}_{h^*}) + 2\xi(n) \leq 3\xi(n).$$

   Therefore, we get that

$$\epsilon_{\mathcal{A}_S}(\mathcal{D}_S, m) = \sup_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{T \sim \mathcal{D}_{h^*}^m}[L(\mathcal{A}_S(T), \mathcal{D}_{h^*})] \leq 5\xi(n) \leq \mathop{\mathbb{E}}_{T \sim \mathcal{D}^m}[\mathcal{C}_S(T)] \leq 7\xi(n) = \epsilon_n(\mathcal{D}_S, m).$$

2. $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_S) \leq \xi(n)$. In this case, we have no guarantees on the success of the uniformity tester. Nevertheless, we show that the error of the certifier can bound the error of the learner. We consider the two possibilities, namely, (1) $h^* = h_S$ and (2) $h^* = h_{S'}$ for $S' \in \mathcal{S}, S' \neq S$.

   In the first possibility, where $h^* = h_S$, we know h_S will have zero error under $\mathcal{D}_{h^*}$ and obviously $L(h_S, T_2) = L(h_S, \mathcal{D}_{h^*}) = 0$.

   Denote $p_1 := \mathcal{D}[\{u \in \mathcal{U}(n) : h^*(u) = 1\}]$ and note that because $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_S) \leq \xi(n)$ we have

$$|\mathcal{D}_S[\{u \in \mathcal{U}(n) : h^*(u) = 1\}] - p_1| \leq \xi(n).$$

   In the case where $h^* = h_{S'}$ for some $S' \neq S$, we know $\mathcal{D}_S[\{u \in \mathcal{U}(n) : h^*(u) = 1\}] \geq 1 - n^{-\beta/2}$ and therefore $\mathbb{E}_{T_1}[r] = p_1 \cdot |T_1| \geq (1 - 2n^{-\beta/2})|T_1|$. A Chernoff bound, for sufficiently large n, concludes that

$$\begin{aligned}
\mathop{\mathbb{P}}_{T}\left[r \leq \frac{1}{2} \cdot |T_1|\right] &= \mathop{\mathbb{P}}_{T}\left[r \leq \frac{1}{2(1-2n^{-\beta/2})} \cdot (1 - 2n^{-\beta/2})|T_1|\right] \\
&\leq \mathop{\mathbb{P}}_{T}\left[r \leq (1 - \frac{1}{3}) \cdot (1 - 2n^{-\beta/2})|T_1|\right] && \left(\frac{1}{2(1-2n^{-\beta/2})} \leq \frac{2}{3}\right) \\
&\leq \exp\left(-\frac{1}{18}(1 - 2n^{-\beta/2})|T_1|\right) \\
&\leq \exp\left(-\frac{m}{144}\right) \leq \xi(n) && \left((1 - 2n^{-\beta/2}) \geq \frac{1}{2}\right).
\end{aligned}$$

   This implies that with probability at least $1-\xi(n)$, we have $r > |T_1|/2$ and $\mathcal{A}_{\mathsf{Maj}}(T_1)(x) = 1$ for all x. Therefore, $L(\mathcal{A}_{\mathsf{Maj}}(T_1), \mathcal{D}_{h^*}) \leq 1 - \mathcal{D}[S \setminus S'] \leq 2\xi(n)$, where the first inequality is due to majority being correct on $S \setminus S'$ and the second inequality is due to $\mathcal{D}[S \setminus S'] \geq \mathcal{D}_S[S \setminus S'] - \xi(n) \geq 1 - 2\xi(n)$. Again, we proved that with probability at least $1 - \xi(n)$ we have $L(\widehat{h}, \mathcal{D}_{h^*}) \leq 2\xi(n)$ for $\widehat{h} \in \arg\min_{h \in \{h_S, \mathcal{A}_{\mathsf{Maj}}(T_1)\}} L(h, \mathcal{D}_{h^*})$. Similar to the previous case, we can take a union bound over the above and the failure of Equation (11) to conclude that with probability at least $1-2\xi(n)$ over $T \sim \mathcal{D}^m$ we have $L(\mathcal{A}_S(T), \mathcal{D}_{h^*}) \leq L(\widehat{h}, \mathcal{D}_{h^*}) + 2\xi(n) \leq 4\xi(n)$. This proves that

$$\epsilon_{\mathcal{A}_S}(\mathcal{D}, m) = \sup_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{T \sim \mathcal{D}^m_{h^*}}[L(\mathcal{A}_S(T), \mathcal{D}_{h^*})] \leq 6\xi(n) \leq \mathop{\mathbb{E}}_{T \sim \mathcal{D}^m}[\mathcal{C}_S(T)] \leq \epsilon_n(\mathcal{D}, m) = 1.$$

3. $\mathsf{d_{TV}}(\mathcal{D}, \mathcal{D}_S) > \xi(n)$. In this case we know that with probability at least $1-\xi(n)$ over $T \sim \mathcal{D}^m$ the test $\mathsf{MTestUnif}_S(\xi(n), \xi(n), T)$ rejects $\mathcal{D}$ and outputs 0. Therefore, $\mathbb{E}_{T \sim \mathcal{D}^m}[\mathcal{C}_S(T)] \geq 1 - \xi(n)$.

   Similar to the reasoning in the Case 3 of the proof of Lemma C.1, we can conclude that for any distribution $\mathcal{D}$, the error of the majority learner is always upper bounded by $3/4$ for large n. We get from Equation (11) that with probability at least $1 - \xi(n)$ over $T \sim \mathcal{D}^m$,

$$L(\mathcal{A}_S(T), \mathcal{D}_{h^*}) \leq L(\mathcal{A}_S(T), T_2) + \xi(n) \leq L(\mathcal{A}_{\mathsf{Maj}}(T_1), T_2) + \xi(n) \leq L(\mathcal{A}_{\mathsf{Maj}}(T_1), \mathcal{D}_{h^*}) + 2\xi(n).$$

   This concludes that

$$\epsilon_{\mathcal{A}_S}(\mathcal{D}, m) = \sup_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{T \sim \mathcal{D}^m_{h^*}}[L(\mathcal{A}_S(T), \mathcal{D}_{h^*})] \leq \frac{3}{4} + 3\xi(n) \leq \frac{7}{8} \leq \mathop{\mathbb{E}}_{T \sim \mathcal{D}^m}[\mathcal{C}_S(T)] \leq \epsilon_n(\mathcal{D}, m),$$

   where we used the fact that $\xi(n) \leq 1/24$ for large n.

Overall, we proved that $\mathcal{C}_S$ is a sound certifier for $\mathcal{A}_S$ and together the collections $\mathcal{A}_S$ and $\mathcal{C}_S$ for $S \in \mathcal{S}(n)$ witness the certifiable error rate $\epsilon_n(.,.)$. ■

E.4. Proof of Lemma 4.5

Notations. For any function $h : \mathcal{U} \to \{0,1\}$, any $(x,y) \in \mathcal{U} \times \{0,1\}$, define $L(h,(x,y)) := \mathbb{1}[h(x) \neq y]$. Moreover, for any unlabeled set $W \in \mathcal{U}^*$ and labeling function h^*, define $L(h,W,h^*) := \frac{1}{|W|}\sum_{x\in W} L(h,(x,h^*(x)))$. For distribution $\mathcal{D}$ over $\mathcal{U}$, and labeling function h^*, define $L(h,\mathcal{D},h^*) := \mathbb{E}_{x\sim\mathcal{D}}[L(h,(x,h^*(x)))]$. For any multiset $W \in \mathcal{U}^*$, we denote by $\mathsf{supp}(W) \subseteq \mathcal{U}$ the set of all distinct elements in W. For any multiset $W \in \mathcal{U}^*$ and test point x, let $W_x := W \cup \{x\}$. For any multiset $T \in (\mathcal{U}(n) \times \{0,1\})^*$, we denote by $\mathsf{dom}(T) \in \mathcal{U}(n)^*$ the unlabeled part of T. Finally, recall that for any multiset $W \in \mathcal{U}^*$, we denote $\mathcal{S}_{\mathsf{supp}(W)} = \{S \in \mathcal{S}(n) : \mathsf{supp}(W) \subseteq S\}$.

Proof

Let $\mathbb{D}_n$ be the uniform distribution over $\{\widetilde{\mathcal{D}}_S : S \in \mathcal{S}(n)\}$ where $\widetilde{\mathcal{D}}_S := (\mathcal{D}_S, h_S)$ defines a learning instance with $\mathcal{D}_S$, the uniform distribution on S, being the marginal distribution and h_S being the labeling function. We will show that a randomly picked distribution from $\mathbb{D}_n$ is expected to incur high error on $\mathcal{A}$. Formally, we will prove that

$$\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}_S\sim\mathbb{D}_n}\mathop{\mathbb{E}}_{T\sim\widetilde{\mathcal{D}}_S^m}[L(\mathcal{A}(T),\mathcal{D}_S,h_S)] = \mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}_S\sim\mathbb{D}_n}\mathop{\mathbb{E}}_{T\sim\widetilde{\mathcal{D}}_S^m}[L(\mathcal{A}(T),S,h_S)] \geq \frac{1}{2} - 2n^{-\beta}, \tag{12}$$

this will be enough to show the existence of the claimed set $S^*_{n,m}$.

For any multiset T, denote $\overline{T} := \mathsf{dom}(T)$. Observe that drawing $T \sim \widetilde{\mathcal{D}}_S^m$ is equivalent to drawing the unlabeled multiset $\overline{T} \sim \mathcal{D}_S^m$ and then labeling it with h_S. Moreover, for any $S \in \mathcal{S}(n)$ and any $W \in S^m$ we have $\mathcal{D}_S^m[W] = |S|^{-m} = n^{-m}$. Therefore, we can write that

$$\begin{aligned}\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}_S\sim\mathbb{D}_n}\mathop{\mathbb{E}}_{T\sim\widetilde{\mathcal{D}}_S^m}[L(\mathcal{A}(T),\mathcal{D}_S,h_S)] &= \mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}_S\sim\mathbb{D}_n}\mathop{\mathbb{E}}_{T\sim\widetilde{\mathcal{D}}_S^m}\left[\frac{1}{n}\sum_{x\in S} L(\mathcal{A}(T),(x,h_S(x)))\right] \\ &= \frac{1}{|\mathcal{S}(n)|}\sum_{S\in\mathcal{S}(n)}\left[\frac{1}{n^m}\sum_{\overline{T}\in S^m}\left[\frac{1}{n}\sum_{x\in S} L(\mathcal{A}(T),(x,h_S(x)))\right]\right].\end{aligned} \tag{13}$$

Fix any $T \in (\mathcal{U}(n)\times\{0,1\})^m$ and $x \in \mathcal{U}(n)$ such that $x \notin \overline{T}$. We do not need to consider any T with $\overline{T} \not\subseteq \mathcal{U}(n)$ since they have zero probability and do not contribute to Equation (13).

Since $|\overline{T}| = m \leq n^{1-\beta}$, we know from property (ii) of $\mathcal{H}(n)$ in Lemma 4.3 that for any labeling $b \in \{0,1\}^{\mathsf{supp}(\overline{T}_x)}$,

$$(1-n^{-\beta})\frac{|\mathcal{S}_{\mathsf{supp}(\overline{T}_x)}|}{2^{|\mathsf{supp}(\overline{T}_x)|}} \leq |\{h_S \in \mathcal{H}(n) : S \in \mathcal{S}_{\mathsf{supp}(\overline{T}_x)}, h_S|\mathsf{supp}(\overline{T}_x) = b\}| \leq (1+n^{-\beta})\frac{|\mathcal{S}_{\mathsf{supp}(\overline{T}_x)}|}{2^{|\mathsf{supp}(\overline{T}_x)|}}.$$

Moreover, it is obvious that for any S, S', we have $h_S|\mathsf{supp}(\overline{T}_x) = h_{S'}|\mathsf{supp}(\overline{T}_x)$ if and only if $h_S|\overline{T}_x = h_{S'}|\overline{T}_x$ as $\mathsf{supp}(\overline{T})$ is the set of unique elements in $\overline{T}$. Therefore, for any labeling $\ell(x)$ of x, there exists a unique labeling $b \in \{0,1\}^{\mathsf{supp}(\overline{T}_x)}$ such that $h_S|\overline{T}_x = (T,(x,\ell(x)))$ if and only if $h_S|\mathsf{supp}(\overline{T}_x) = b$. This combined with the above equation implies that for any T and $x \notin T$, the

fraction of sets S that contain $\mathsf{supp}(\overline{T}_x)$, are consistent with the labeling of T, and label x with 1 is close to the fraction that label x as 0. Formally, we have

$$\min_{y\in\{0,1\}} \frac{\left|\{h_S \in \mathcal{H}(n) : S \in \mathcal{S}_{\mathsf{supp}(\overline{T}_x)}, h_S|\overline{T} = T, h_S(x) = y\}\right|}{\left|\{h_S \in \mathcal{H}(n) : S \in \mathcal{S}_{\mathsf{supp}(\overline{T}_x)}, h_S|\overline{T} = T\}\right|} \geq \frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}.$$

Define by $\mathcal{S}[(T,x)] = \{S \in \mathcal{S}(n) : S \in \mathcal{S}_{\mathsf{supp}(\overline{T}_x)}, h_S|\overline{T} = T\}$ the collection of sets that contain $\mathsf{supp}(\overline{T}_x)$ and their labeling function is consistent with the labels of T. The above implies that for any learner $\mathcal{A}$ and for any set T and $x \notin T$ we have

$$\frac{1}{|\mathcal{S}[(T,x)]|}\sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x))) \geq \frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}. \tag{14}$$

Note that for any $S \in \mathcal{S}(n)$, the multiset T and $x \notin T$ have non-zero probability under $\mathcal{D}_S$ as training and test samples if and only if $\mathsf{supp}(\overline{T}_x) \subseteq S$ (i.e., $S \in \mathcal{S}_{\mathsf{supp}(\overline{T}_x)}$), and $h_S|\overline{T} = T$. Therefore, for any T and $x \notin T$, the collection $\mathcal{S}[(T,x)]$ are exactly all the sets $S \in \mathcal{S}(n)$ for which T and x have non-zero probability as training and test samples from $\mathcal{D}_S$. In other words, each set T and test point $x \notin T$ appear as a summand in Equation (13) exactly $|\mathcal{S}[(T,x)]|$ many times. Therefore, we can continue writing

$$\begin{aligned}&\frac{1}{|\mathcal{S}(n)|}\sum_{S\in\mathcal{S}(n)}\left[\frac{1}{n^m}\sum_{\overline{T}\in S^m}\left[\frac{1}{n}\sum_{x\in S} L(\mathcal{A}(T),(x,h_S(x)))\right]\right]\\ &= \frac{1}{|\mathcal{S}(n)|}\cdot\frac{1}{n^{m+1}}\sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n)}}\sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x))).\end{aligned} \tag{15}$$

Now note that from Equation (14) we get that

$$\begin{aligned}&\sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n)}}\sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x)))\\ &= \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n),\, x\notin\overline{T}}}\sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x))) + \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n),\, x\in\overline{T}}}\sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x)))\\ &\geq \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n),\, x\notin\overline{T}}} \frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}\cdot|\mathcal{S}[(T,x)]|.\end{aligned} \tag{16}$$

Moreover, for any fixed $S \in \mathcal{S}(n)$, and any T with $\mathsf{supp}(\overline{T}) \subset S$ there are at most $|T|$ many $x \in \mathcal{U}(n)$ such that $x \in S$ and $x \in \overline{T}$. In other words, there are at least $n - |T| \geq n - m$ many x such that $x \in S$ but $x \notin \overline{T}$. Therefore,

$$\forall S \in \mathcal{S}(n),\quad \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n),\, x\notin\overline{T}}} \mathbb{1}[S \in \mathcal{S}[(T,x)]] \geq \frac{n-m}{n}\sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m\\ x\in\mathcal{U}(n)}} \mathbb{1}[S \in \mathcal{S}[(T,x)]]$$

Observe that this further means

$$\sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n),\, x\notin\overline{T}}} |\mathcal{S}[(T,x)]| \geq \frac{n-m}{n} \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n)}} |\mathcal{S}[(T,x)]| \tag{17}$$

Taking Equations (16) and (17) into account, we can continue Equation (15) to write

$$\begin{aligned}
&\frac{1}{|\mathcal{S}(n)|}\cdot\frac{1}{n^{m+1}} \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n)}} \sum_{S\in\mathcal{S}[(T,x)]} L(\mathcal{A}(T),(x,h_S(x))) \\
&\geq \frac{1}{|\mathcal{S}(n)|}\cdot\frac{1}{n^{m+1}}\cdot\frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}\cdot \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n),\, x\notin\overline{T}}} |\mathcal{S}[(T,x)]| \\
&\geq \frac{1}{|\mathcal{S}(n)|}\cdot\frac{1}{n^{m+1}}\cdot\frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}\cdot\frac{n-m}{n}\cdot \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n)}} |\mathcal{S}[(T,x)]|
\end{aligned}$$

Combining the above with the fact that $\sum_{T,x}|\mathcal{S}[(T,x)]| = \sum_{S\in\mathcal{S}(n)}\sum_{T,x}\mathbb{1}[S\in\mathcal{S}[(T,x)]] = |\mathcal{S}(n)|\cdot n^{m+1}$, we get

$$\begin{aligned}
\frac{1}{|\mathcal{S}(n)|}\cdot\frac{1}{n^{m+1}}\cdot\frac{1}{2}\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}}\cdot\frac{n-m}{n}\cdot \sum_{\substack{T\in(\mathcal{U}(n)\times\{0,1\})^m \\ x\in\mathcal{U}(n)}} |\mathcal{S}[(T,x)]| &\geq \frac{1}{2}\cdot\left(\frac{n-n^{1-\beta}}{n}\right)\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}} \\
&= \frac{1}{2}\cdot\left(1-n^{-\beta}\right)\cdot\frac{1-n^{-\beta}}{1+n^{-\beta}} \\
&\geq \frac{1}{2}-2n^{-\beta}.
\end{aligned}$$

where we used the fact that $m \leq n^{1-\beta}$. This proves that for any fixed deterministic learner and sample size $m \leq n^{1-\beta}$, the expectation over $\widetilde{\mathcal{D}}_S \sim \mathbb{D}_n$ of the error of $\mathcal{A}$ on $\widetilde{\mathcal{D}}_S$ when trained on samples of size m is more than $1/2 - 2n^{-\beta}$. This is enough to show that for any (randomized) learner $\mathcal{A}$ and sample size $m \leq n^{1-\beta}$, there exists a set $S^*_{n,m} \in \mathcal{S}(n)$ with distribution $\mathcal{D}_{S^*_{n,m}}$ such that $\mathcal{A}$ has error at least $1/2 - 2n^{-\beta}$ on $\mathcal{D}_{S^*_{n,m}}$ given samples of size m. ■

Appendix F. Missing Proofs from Section 5

F.1. Proof of Theorem 5.2

We show that any distribution family which is not smartly learnable (in the non-relative sense) can be modified to satisfy the theorem. Such a countable distribution family $\mathbb{D}$ exists for a hypothesis class $\mathcal{H}$ on a countable domain $\mathcal{X}$, as shown in (Darnstädt et al., 2013, Theorem 2).

Define a new domain $\mathcal{X}'$ which "tags" each $x \in \mathcal{X}$ with the name of a distribution $\mathcal{D} \in \mathbb{D}$; formally, $\mathcal{X}' = \mathcal{X} \times \mathbb{D}$. Now extend $\mathcal{H}$ to $\mathcal{X}'$ by simply ignoring the tags; i.e., for each $h \in \mathcal{H}$ define $h'(x,\mathcal{D}) = h(x)$ and $\mathcal{H}' = \{h' : h \in \mathcal{H}\}$. Finally, port each distribution $\mathcal{D} \in \mathbb{D}$ to the subset of the domain tagged with $\mathcal{D}$, preserving probabilities; i.e., let $\mathcal{D}'$ be such that $\mathbb{P}_{\mathcal{D}'}[(x,\mathcal{D})] = \mathbb{P}_{\mathcal{D}}[x]$, and let $\mathbb{D}' = \{\mathcal{D}' : \mathcal{D} \in \mathbb{D}\}$. Clearly $\mathcal{X}'$ and $\mathbb{D}'$ are countable.

Any sample from a distribution in $\mathcal{D}' \in \mathbb{D}'$ uniquely identifies it, so the learner which identifies $\mathcal{D}'$ then implements the optimal distribution-fixed learner for it is $(1, m)$-smart. For the same reason, the optimal distribution-fixed error for each $\mathcal{D}' \in \mathbb{D}'$, equal to the optimal distribution-fixed error for its precursor $\mathcal{D} \in \mathbb{D}$, is soundly certifiable. Any learner which ignores the tags, such as ERM or OIG, does no better on $\mathcal{D}'$ than on its precursor $\mathcal{D}$, for any sample size. Recalling that $\mathbb{D}$ is not smartly learnable, we conclude that no such learner is relatively smart over $\mathbb{D}'$. ■